Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/73/20240411

DimSum50: A benchmark dataset of Chinese food

Kunyi Yu

Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China

12013027@mail.sustech.edu.cn

Abstract. This paper provides an overview of the increasing attention given by the public and government to food health. It also reviews the progress made in the field of image classification over the past decade, with a specific focus on the current state of Chinese food datasets. The DimSum50 dataset is introduced as the first dataset that concentrates on diverse Chinese dim sum among all publicly available datasets. This dataset comprises 50 categories of the most popular dim sum foods, containing a total of 28,884 images. To ensure the accuracy and scalability of DimSum50, a three-step construction process was implemented, including category selection, data collection, and data cleaning. The unique properties of dim sum present challenges in constructing this dataset, as several categories exhibit similar characteristics. Benchmark experiments were conducted on the DimSum50 dataset, offering a horizontal comparison among several common and state-of-the-art models in both CNNs and transformers.

Keywords: Chinese food classification, Benchmark dataset, DimSum50, Machine Learning, Deep Learning.

1. Introduction

Health is a topic of significant importance today, as people prioritize their well-being and defense against illnesses. According to the United Nations study "World Population Prospects 2022" [1], the average life expectancy worldwide has increased from 46.5 years in 1950 to 71.0 years in 2021. Additionally, the United States' National Health Expenditures have seen a significant rise, from 27.1 million dollars in 1960 to 4,255.1 million dollars in 2021 [2]. Correspondingly, Personal Health Care Expenditures for U.S. citizens have increased from 62.4 dollars in 1970 to 3,553.4 dollars in 2021 [3].

Researchers from both industry and academia have also dedicated considerable attention to studying health, resulting in numerous related research endeavors. Food, being the primary energy and nutrient source for human beings, is a particularly significant area of research within the field of health. Food analysis, including the measurement of chemical composition, effectively assists ordinary individuals in understanding their overall nutritional intake and optimizing their diets. This is crucial for maintaining a balanced diet, preventing diseases, and promoting good health. Calorie analysis plays a vital role in diet management and health maintenance, enabling individuals to better control their intake and avoid overeating or energy deficiencies.

Food type classification is an emerging technique within the field of food analysis, as it allows for the automatic evaluation of food on a large scale. The development of food type classification techniques is based on image classification, a more generalized field within computer vision. Considerable efforts

DOI: 10.54254/2755-2721/73/20240411

have been dedicated to image classification, involving the proposal of better models or algorithms and the establishment of larger image datasets.

The first landmark in image classification was the introduction of AlexNet, a conventional neural network (CNN) model that achieved high accuracy. AlexNet won the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) [4], achieving a top-5 error rate of 15.3%, more than 10.8% lower than the second-best entry. Although not the first GPU-implemented CNN model, AlexNet's innovative techniques, such as deep architecture, ReLU activation function, and dropout regularization, made CNN an attractive research field and stimulated further creative work. AlexNet demonstrated the robustness and potential of CNN-based classifiers, serving as the foundation for modern deep CNN models [4].

Subsequently, other landmark CNN models were proposed. VGGNet and GoogLeNet were the runner-up and winner of ILSVRC2014, respectively. VGGNet's main contribution was increasing depth using a few large kernel-size filters, followed by multiple very small convolution filters of size (3x3). VGGNet successfully constructed CNN models with up to 19 layers. GoogLeNet, on the other hand, is a 22 layers CNN model which also known as the first version of the Inception network. GoogLeNet addressed overfitting in deep CNNs by using filters with multiple sizes between adjacent layers, increasing the model's width rather than depth. However, training deeper CNN models remained challenging due to the common issue of vanishing/exploding gradients.

ResNet (Residual Network) introduced a new framework to alleviate the training process of deeper networks. Early models often faced a degradation problem, where the accuracy of the model rapidly decreased as the number of layers increased. ResNet addressed this issue by adding shortcut connections that enabled layers to fit a residual mapping. He et al. (2016) successfully demonstrated that this approach allowed the network to be seven times larger than VGG in terms of network layers, with lower time complexity and improved image classification performance.

The release of the Transformer model in 2017 [5] brought astonishing results in the natural language processing community, as it effectively handled sequential elements. The Vision Transformer (ViT) [6] was the first model based solely on the Transformer model. After extensive data pre-training and fine-tuning processes, the ViT model exhibited excellent results in image classification tasks compared to state-of-the-art CNNs, with lower energy costs. Additionally, the DeiT [7] achieved successful training on the relatively smaller ImageNet database by combining attention-based distillation, using CNNs as teachers and transformers as students, resulting in competitive results with fewer parameters [8].

In recent years, with the rapid development of mobile and terminal devices, greater attention has been directed towards areas with limited computing resources. Several lightweights and efficient milestone CNN models have been proposed, such as MobileNet, ShuffleNet, and EfficientNet. MobileNet employs depthwise separable convolutions and two simple global hyperparameters to strike a balance between resource usage and accuracy. ShuffleNet architecture utilizes pointwise group convolution and channel shuffling, enabling it to outperform MobileNet with a computation limitation of 40 MFLOPs and achieve a 13 times speedup over AlexNet. EfficientNet, on the other hand, utilizes the compound scaling method to balance accuracy and efficiency, a technique that can be implemented in other models such as MobileNet and ResNet.

The establishment of several landmark datasets, such as MNIST, CIFAR-10/100, and ImageNet, has significantly propelled the development of image classification. Traditional and small-scale datasets, such as MNIST and CIFAR-10/100, provided foundational platforms and benchmarks for training, testing, and fine-tuning models before 2009. However, these datasets had limitations in terms of diversity, complexity, and scaling models. The introduction of the "ImageNet" dataset, initiated by Feifei Li in 2009, revolutionized the field of computer vision by providing a vast and diverse dataset with millions of high-resolution images. The ImageNet competitions, held eight times from 2010 to 2017, consistently produced landmark models such as AlexNet, ResNet, and VGG. The development of computer vision models owes much to the availability of high-quality image data.

Nevertheless, food image datasets still fall short compared to general image datasets. As of Winter 2021, ImageNet already includes over 21,841 categories with 14,197,122 well-labeled images. In

DOI: 10.54254/2755-2721/73/20240411

contrast, the data scale of food image datasets remains much smaller. Therefore, one of the main objectives is to advance this area and establish Chinese dim sum datasets. Food image datasets encompass various categories of images with similar shapes and colors, presenting greater challenges within the field of image classification.

To meet these requirements, the DimSum50 dataset is proposed. This work involves obtaining a high-quality dataset through a three-step process that includes meticulous manual screening. The proposed dataset aims to enhance the demand for automated food health control among the public and accelerate the development of the field of food image classification. The rest of this paper is organized as follows: Section II discusses existing food datasets, Section III introduces the food dataset proposed in this paper, Section IV provides benchmark experiments, and Section V discusses the challenges in this area and future work.

2. Existing Food Image datasets

This section provides an overview of existing food image datasets, including early attempts, widely used benchmark datasets, and Chinese food datasets. These datasets have been instrumental in the advancement of research in this field.

Table 1. Datasets mentioned in this section.

Dataset	Present in	Classes	Images per class	Images
PFID	2009	101	45	4,545
UEC Food-100	2012	100	90.6	9,060
UEC Food-256	2014	256	≈112.6	31,395
ETHZ Food-101	2014	101	1,000	101,000
ChineseFoodNet	2017	208	892.4	185,628
ChinFood1000	2017	1000	less than 120	NA
CF-108	2020	108	≈93.3	100,800

The establishment of food image datasets started around the same time as general image datasets. One of the first datasets in this domain was the Pittsburgh fast-food image dataset (PFID), introduced in 2009 by Yanai and Joutou [9]. PFID contains 4,545 still images belonging to 101 semantically meaningful categories. It includes images from various fast-food chains and includes stereo data and videos. Although PFID established benchmarks using two approaches, the accuracy was not satisfactory due to technological limitations. Nevertheless, PFID served as a groundbreaking food dataset that inspired further research in this area.

In the 2010s, several significant benchmark datasets emerged, accompanied by the development of classifiers and technologies. These datasets include:

UEC Food-100 [10]: Released in 2012, UEC Food-100 focuses on Japanese foods instead of American fast foods. It comprises 9,060 pictures with 100 categories, with at least 500 of them being multi-item food. The inclusion of multi-item food enhances the difficulty of food identification.

UEC Food-256 [11]: Proposed by Yoshiyuki and Keiji in 2014, UEC Food-256 introduces a new transfer framework that combines existing categories, support vector machines (SVMs), and crowdsourcing to release a new dataset with more than 256 kinds of food from various countries, including French, Italian, and American cuisines. It contains 31,395 images, and each image is accompanied by a bounding box that indicates the location of the food. This dataset is particularly suitable for tasks such as localization or object detection in addition to image classification.

ETHZ Food-101 [12]: ETHZ Food-101 is a real-world food dataset collected from a website where users can upload their dietary information and food types. This dataset intentionally includes some uncleaned noise in the training images to challenge models to exhibit better robustness. It contains 101 food categories.

DOI: 10.54254/2755-2721/73/20240411

Chinese cuisine is known for its diverse range of flavors, appearances, and regional specialties. Consequently, there are several relevant image datasets specifically focusing on Chinese food. These datasets include:

ChineseFoodNet [13]: ChineseFoodNet is a dataset collected from recipe pictures or selfies, consisting of 208 food categories and over 180,000 images. It covers a wide variety of popular Chinese dishes.

ChinFood1000 [14]: ChinFood1000 contains 1,000 classes, with each class ranging from 30 to 120 images. Over 85% of the classes represent different dishes. Due to the relatively small number of images per class, some deep CNN models may experience underfitting, resulting in a top-5 accuracy of less than 70%.

CF-108 [15]: CF-108 is a relatively new Chinese food dataset collected from publicly available images on websites. It comprises 108 food categories and includes 100,800 images. The authors have conducted data cleaning, smoothing, and labeling to improve the quality of the dataset.

3. DimSum50 dataset

The existing general food dataset already fulfills some needs of relevant research due to its sufficient data amount and rich diversity. However, Chinese food datasets are still inadequate compared to other food types. While some of them contain Chinese dishes ranging from north to south, there is still a significant gap in datasets specifically focused on snacks and dim sum images. Eating dim sum is a popular dietary habit in China, with a strong mass base and historical heritage. Different types of snacks vary greatly, and there are also many types that appear very similar, posing a significant challenge to the construction of datasets and image recognition.

3.1. Category Selection

In each Chinese cuisine, such as Cantonese food, Sichuan food, and Beijing food, there are several famous dim sum dishes. The research team initially selects representative dim sum dishes that are popular on social media platforms and downloads their typical pictures. Over 60 types of snacks are considered as candidates. However, some of them have visual similarities and cannot be easily differentiated by humans, such as two types of Chicken Feet, Double Skin Milk, and Ginger Milk Curd. To address this issue, several categories are combined into composite categories. As a result, 50 categories of Chinese dim sum are chosen to create the DimSum50 dataset. Table 2 and figure 1 provide an overview of the categories and sample images included in the dataset.

Table 2. Names of the 50 categories in DimSum50

Naan	Mahua	Mantou	Zongzi	Youtiao
Huajuan	Shaomai	Roujiamo	Tangyuan	Mooncake
Egg Tart	Lyudagunr	Ice Jelly	Fried Pork	Egg Waffle
Fried Milk	Frozen Pear	Rice Congee	Turnip Cake	Bingtanghulu
Spring Rolls	Sponge Cake	Walnut Baozi	Chicken Feet	Egg Yolk Puff
Guiling Jelly	Jianbing guozi	Pineapple Baozi	Roasted Chestnuts	Steamed Rice Rolls
Sweet Potato Balls	Pan-Fried Raviolis	Jiaozi with Shrimp	Marinated Beef Tripe	Stewed Beef Meatballs
Steamed Red Rice Rolls	Steamed Brown Sugar Cake	Rice Noodle in Clear Soup	Baozi Stuffed with Custard	Baozi Stuffed with BBQ Pork
Rice Pudding with Brown Sugar	Steamed and Deep-Fried Mantou	Double Skin Milk&Ginger Milk Curd	Stir-Fried Rice Noodles with Beef	Pan-Fried Baozi Stuffed with Pork
Wonton Soup in Hot and Spicy Sauce	Steamed Chicken with Glutinous Rice	Steamed Spare Ribs in Black Bean Sauce	Noodles with Soy Bean Paste, Beijing Style	Sichuan Steamed Pork Buns Wrapped in Leaves

DOI: 10.54254/2755-2721/73/20240411

Figure 1. Representative images the 50 categories in DimSum50

3.2. Data collection

The pictures used in the dataset are obtained from two sources. The majority of the images are crawled from the Baidu website [1], which contains a large number of food images from the Chinese Internet. In this step, more than 40,000 images are downloaded. Each category has more than 400 pictures, with the majority having around 1,000 pictures. However, it is expected that many of these images are misclassified or contain advertising texts or large watermarks, so the dataset requires several steps of data cleaning.

3.3. Data cleaning

After collecting a large number of images, data cleaning is necessary to ensure the accuracy and usability of the dataset. This process is time-consuming and involves four stages, as shown in figure 2: removing images with unbalanced sizes, converting images to a uniform storage format (jpg), removing misclassified images, and cropping images.

Firstly, images with abnormal length-width ratios outside the range of 0.5 to 2 are removed. This is because resizing images to squares during pre-processing can cause abnormal compression distortions. The second step involves converting the images to the jpg format, which is frequently used and can preserve the original visual information in the three RGB channels. The last two steps, removing misclassified images and cropping images, are the core parts of the data cleaning process. Due to the time-consuming nature of these processes, several volunteers were recruited to assist in the data cleaning. The dataset removes images containing highly occluded advertising texts, while some images with small and unnoticed watermarks are retained. The final step, cropping images, aims to centralize the dim sum objects and ensure that the images contain more useful information about the dim sum itself, rather than distracting backgrounds.

[1] Images.baidu.com

DOI: 10.54254/2755-2721/73/20240411

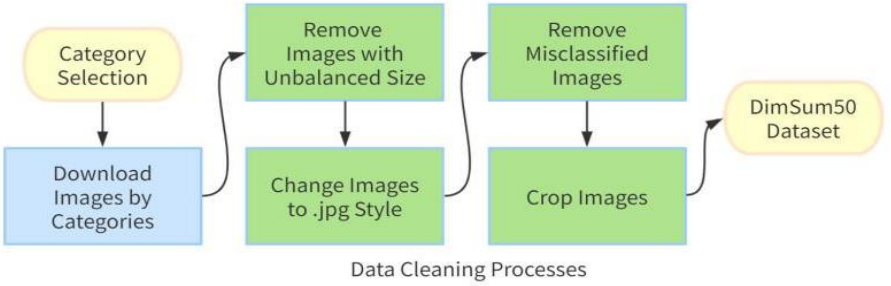

Figure 2. The process of establishing DimSum50

3.4. Dataset Description

After completing the three-step dataset construction mentioned in the previous subsections, the DimSum50 dataset contains 50 categories, 28,884 images, and has a size of approximately 2.11 gigabytes. All the images in the dataset retain their original colors, which may include directly photographed images and edited images posted on the internet. The DimSum50 dataset ensures that the size of the images is not smaller than 256 x 256 pixels. Since dim sum is usually made and arranged together, a single picture often contains multiple dim sum objects of the same type.

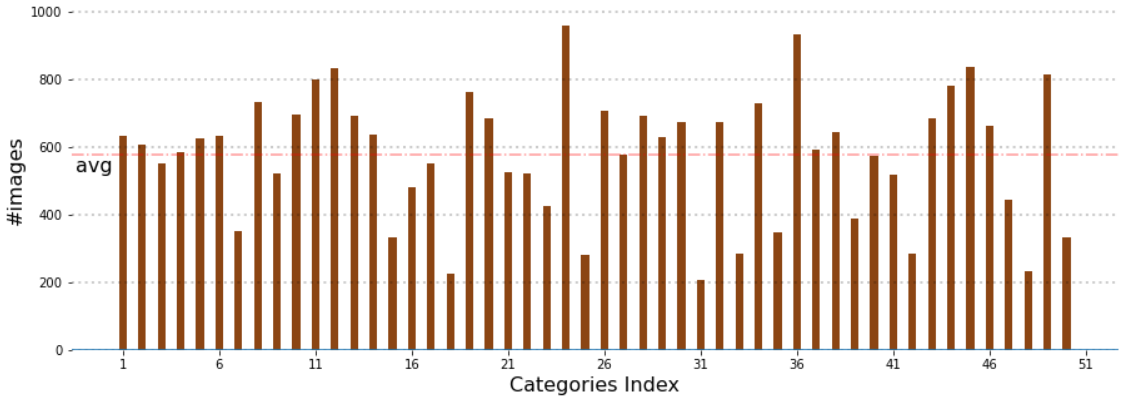

Figure 3. Number of images per category

Figure 3 illustrates the distribution of the number of images in each category. The number of images ranges from more than 200 to nearly 1,000 images, resulting in a balanced distribution. The average number of images in each category is 577.68, with a median of 615. The DimSum50 dataset is divided into three parts: training, validation, and testing sets, approximately at a ratio of 8:1:1. Specifically, there are 23,106, 2,889, and 2,889 images in the training, validation, and testing sets, respectively.

The DimSum50 dataset and relevant information are available for non-commercial use on the website: sites.google.com/view/dimsum50.

4. Benchmark Experiments

In this section, benchmark experiments with the DimSum50 dataset using different CNN models and transformer models are discussed. Three metrics are recorded in the experiments: Top-1 accuracy, Top-5 accuracy, and average Top-1 accuracy of all models, which can provide a comprehensive insight into the properties of the DimSum50 dataset and the performances of the models.

DOI: 10.54254/2755-2721/73/20240411

Table 3. Experiment results on DimSum50 dataset.

Models	Top-1 Accuracy	Top-5 Accuracy
VGG11	0.80478	0.95639
VGG19[a]	0.80824	0.96919
ResNet18	0.85289	0.97092
ResNet50	0.87712	0.98339
ResNet152	0.88231	0.98200
EfficientNet	0.89547	0.98719
MobileNet	0.87574	0.98442
ShuffleNet	0.85843	0.97404
VisionTransformer[a]	0.91381	0.99065
SwinTransformer[a]	**0.93008**	**0.99342**

a: using frozen operation

The results of the comparison between the different models are shown in table 3. VGG and ResNet are labeled with specific model configurations. The other 4 models were tested in their up-to-date and smallest version of the Python torchvision package, namely efficientnet_v2_s, mobilenet_v3_small, vit_b_16, and swin_v2_t, respectively. All models in the experiments are trained with 50 epochs and all use the latest pre-trained parameters in the first epoch of training. Moreover, the part of the models using the frozen technique means that they only update the last linear output layer, the classifier in CNN or the head in transformers, leaving the previous parameters unchanged. These partial models training without frozen operations show very poor results, perhaps due to problems with model scaling and the training method. After using freezing, both the training speed and classification performance are improved tremendously.

The results show that the Top-1 accuracy of the two Transformer models on the DimSum50 dataset is over 90%, which is a significant improvement over the 80%+ accuracy of the other CNN models in this test. For the Top-5 accuracy, all models achieve more than 95%. The experiment shows that the Transformer-based models are already taking computer vision to new heights.

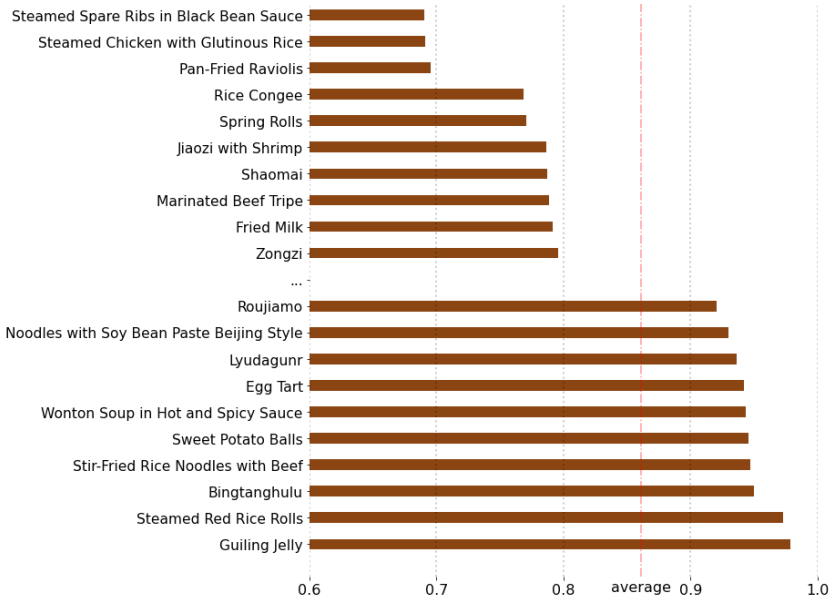

Figure 4. The 10 most difficult and the easiest categories averaged across all approaches.

DOI: 10.54254/2755-2721/73/20240411

The result of the experiment also analyzes the average Top-1 accuracy of all models by category, as shown in figure 4. The 10 most difficult categories are at the top of the figure, and the 10 easiest categories are at the bottom. The most difficult category, steam chicken with glutinous rice, has a variety of shapes at different stages of preparation and consumption, and some steps are similar to Zongzi. In addition, the appearance of spring rolls and fried milk is extremely similar even for the uninitiated. This shows that the selection of dim sum types is quite a challenge. However, it may also be due to the fact that there are fewer images for certain classes, and this may be a point where the dataset can be improved in the future. However, the 10 simplest categories also show that the overall accuracy of the dataset is very high as all image types are selected through the same process.

5. Conclusions

This paper reviews the background of public attention to food health and recent developments in image recognition over the last decade. The main contribution of this paper is the introduction of DimSum50, a dataset consisting of 50 categories of images of Chinese dim sum dishes. Due to the high similarity of dim sum dishes in the distinct categories and the presence of multiple objects in the same image, this dataset provides an exciting opportunity to test and advancing visual algorithms. The benchmark experiments also provide results that are suitable for comparison with other models.

As for future work, the dataset could be extended using transfer learning and automated methods instead of relying on manual data cleaning to improve accuracy and usability. In addition, low-power food recognition technology suitable for end devices can be developed based on the dataset.

References

[1] United Nations, Department of Economic and Social Affairs, Population Division 2022 *World Population Prospects 2022, Online Edition*

[2] Centers for Medicare and Medicaid Services. NHE Summary, including share of GDP, CY 1960-2021 (ZIP) [Internet]. Baltimore (MD): CMS; [last updated 2022 Dec 15]. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical

[3] Centers for Medicare and Medicaid Services. NHE Tables (ZIP) [Internet]. Baltimore (MD): CMS; [last updated 2022 Dec 15]. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical

[4] Feng X, Jiang Y, Yang X, Du M and Li X 2019 Computer vision algorithms and hardware implementations: A survey *Integration* 69 309-20

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in neural information processing systems* 30

[6] Dosovitskiy A *et al* 2020 An image is worth 16x16 words: Transformers for image recognition at scale *arXiv preprint arXiv:2010.11929*

[7] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A and Jégou H 2021 Training data-efficient image transformers & distillation through attention *International Conference on Machine Learning* pp 10347-57

[8] Khan S, Naseer M, Hayat M, Zamir SW, Khan FS and Shah M 2022 Transformers in vision: A survey *ACM Computing Surveys (CSUR)* 54(10s) 1-41

[9] Joutou T and Yanai K 2009 A food image recognition system with multiple kernel learning *2009 16th IEEE International Conference on Image Processing (ICIP)* pp 285-8

[10] Matsuda Y, Hoashi H and Yanai K 2012 Recognition of multiple-food images by detecting candidate regions *2012 IEEE International Conference on Multimedia and Expo* pp 25-30

[11] Kawano Y and Yanai K 2014 Automatic expansion of a food image dataset leveraging existing categories with domain adaptation *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III* 13 pp 3-17

DOI: 10.54254/2755-2721/73/20240411

[12] Bossard L, Guillaumin M and Van Gool L 2014 Food-101–mining discriminative components with random forests *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI* 13 446-61
[13] Chen X, Zhu Y, Zhou H, Diao L and Wang D 2017 Chinesefoodnet: A large-scale image dataset for Chinese food recognition *arXiv preprint arXiv:1705.02743*
[14] Fu Z, Chen D and Li H 2017 Chinfood1000: A large benchmark dataset for chinese food recognition Intelligent Computing Theories and Application: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part I 13 pp 273-81
[15] Li Y, Xu X and Yuan C 2020 Enhanced mask r-cnn for chinese food image detection *Mathematical Problems in Engineering* 2020 1-8