


# Context-aware CLIP for Enhanced Food Recognition

Övgü Öztürk Ergün<sup>1</sup> 

<sup>1</sup> Department of Software Engineering, Faculty of Engineering, Muğla Sıtkı Koçman University, Türkiye

## Abstract

Generalization of food image recognition frameworks is difficult due to the wide variety of food categories in cuisines across cultures. The performance of the deep neural network models highly depends on the training dataset. To overcome this problem, we propose to extract context information from images in order to increase the discrimination capacity of networks. In this work, we utilize the CLIP architecture with the automatically derived ingredient context from food images. A list of ingredients are associated with each food category, which is later modeled as text after a voting process and fed to a CLIP architecture together with input image. Experimental results on the Food101 dataset show that this approach significantly improves the model's performance, achieving a 2% overall increase in accuracy. This improvement varies across food classes, with increases ranging from 0.5% to as much as 22%. The proposed framework, CLIP fed with ingredient text, outperforms Yolov8 (81.46%) with 81.80% top 1 overall accuracy over 101 classes.

**Keywords:** Food Image Processing; CLIP; Ingredient Analysis; Deep Learning; Context; AI.

## 1. Introduction

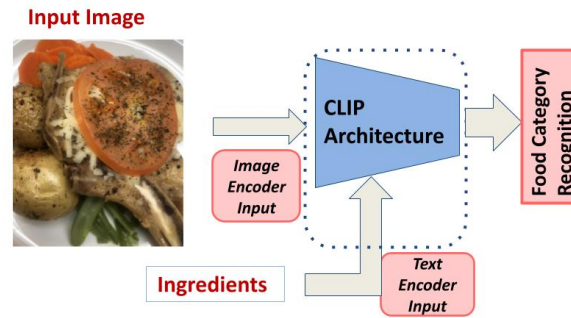
Technological advances in food recognition have greatly benefited from fusion [1-4] of image and text data, leading to improved content retrieval and context understanding. These multi-modal approaches leverage graph networks, entropy-based methods, and statistical learning to obtain meaningful insights that contribute to region of interest identification, ingredient extraction and ontology modeling. The use of complex context modeling can indeed result in high accuracy, but at the cost of increased computational complexity and model interpretability.

On the other hand, single-context modeling approaches aim to simplify the model by focusing on a single primary source of information, such as text. However, these approaches [5-8] still struggle to achieve satisfactory results in the food domain. This challenge arises from high diversity [3, 9] of food images, which exhibit a wide range of variations in ingredient combinations, presentation styles and the nuances of culinary diversity. Although general image recognition [10] architectures such as RCNN, YOLO are commonly employed [3, 11, 12] in food recognition, they limit domain-specific context tailoring and their performance varies across different food datasets. Recently, CLIP [13] has proven to be a breakthrough in combining image and text for enhanced recognition tasks and successfully applied to many different vision problems. In this work, we present a simple yet effective CLIP-based architecture to boost the performance of food category recognition. Our approach stands as a pioneering approach utilizing CLIP in food domain.

We propose a CLIP-based image and text fusion architecture (**Fig.1**) where the first step involves generation of the list of ingredients. These ingredients are then translated into textual descriptions, which are fed into the text encoder of CLIP system. For testing, we utilize ResNet50 and ViT-B/32 image encoders within CLIP. A wide set of experiments are conducted on Food101 dataset [18] with detailed category-based analysis. Our CLIP-based approach outperforms [7, 11, 14, 15] existing techniques, achieving %81.80 accuracy across 101 classes. Additionally, it delivers competitive performance when compared to state-of-the-art complex architectures [1, 3]. This result not only highlights the potential of leveraging CLIP in food recognition but also presents an exciting opportunity to push the boundaries of food domain with simpler modalities.

\*Corresponding author

E-mail address: ovguozturk@mu.edu.tr



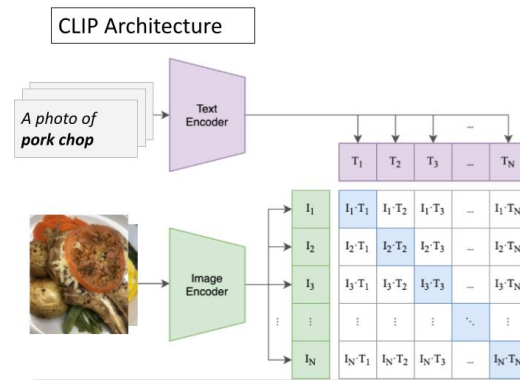
**Figure 1.** Overall Architecture.

## 2. Related Work

In the food domain, there are various research challenges [1, 3, 16], ranging from segmentation tasks to multi label recognition, ingredient extraction, and recipe derivation. Most research study multi-label [19] food recognition, whereas some group tackle food image segmentation [3, 17] problem to distinguish food items or extract portion size. On the other hand, there are solutions focusing on improving food recognition performance via tailored architectures [13], compressed variance [7], and ontology-based [2, 18, 19] semantic cues with limited results. Achieving satisfactory accuracies in food category recognition from single food images still remains an open problem, making it challenging for practical applications. Ponte et al. [2, 19] investigates web scraping and LLM to recognize single and multiple labels in food images in Mexican dataset. Ontology information is driven from text and given to ResNet50 framework together with image, achieving 70% accuracy. Deepfood [19] research reaches 77.4% accuracy on Food101 dataset by employing multi-level DNN architecture. Zhao et al. [4] presents a fusion approach, where class labels are given to BERT architecture and produce context-sensitive embeddings and fed to a few-shot fusion CNN. At the final step, graph convolutional networks are employed for inter-class relation learning. They obtain 68.76% accuracy on Food101 dataset. In their work Mao et al. [11] create a new food image dataset, VIPER-FoodNet (VFN) dataset, consists of food categories with 15k images. They utilize Faster-RCNN for food region localization and present cluster-based proposals to a multi-layered CNN and test their model with different datasets, with 79.81% accuracy in Food101.

Zhang et al. [15] propose a new supervised subnetwork-based feature encoding and pattern classification model generated by a multi-level multiple CNN architectures to leverage feature transformation and fusion efficiently. In [12], Yolov5 based performance evaluation is given across various food datasets in the existence of multiple food items in the images. Min et al. [21] propose extraction of ingredients by first localizing attention region as a reference and sequentially discovering diverse attentional regions with fine-grained scales by means of a multi-level LSTM and spatial transformer networks. The drawbacks of this method are two-fold: the system is high in complexity and highly dependent on region extraction and spatial appearances of ingredients. Chen et al. [22] implement a multi-task DCNN for extracting ingredients from images. However, the number of ingredients is limited and modeled based on food clusters, rather than the basic ingredients that compose the food.

CLIP [13] given in **Fig. 2** has been widely adopted and extended to a variety of vision tasks since its introduction. It leverages the power of large-scale datasets containing both text and images, enabling it to perform well on various vision problems [23] by learning visual representations that are aligned with textual descriptions. Some major applications of CLIP in vision problems include event recognition [24], image/text transformation [20,25], anomaly detection [26] and sketch-based image retrieval [27]. To our knowledge, CLIP architecture has not been yet thoroughly investigated in food recognition problem. The only existing two approaches [28, 29] present poor recognition performances. Very recent study for multi label food image recognition, Rawlekar et al. [28] adapts CLIP to generate prior logits, which is later model by graph CNN representing co-occurrences of different food labels. Results show accuracy %52 in FoodSeg103 dataset. In the other work, Wu et al. [29], very little information is available and it states that fine-tuning of CLIP with only a few samples achieves 63% accuracy on food image recognition, which is a relatively low performance.



**Figure 2.** CLIP Architecture with Image and Text Encoder.

Unlike complex architectures, we present a simple yet high-performance architecture that leverages CLIP to simultaneously incorporate both textual and visual information. Our system achieves significant accuracy improvements across each category, with an overall accuracy of 81.80%, surpassing Yolov8. This highlights the potential of the proposed approach within the food domain, offering significant benefits when further studied for tasks such as meal recognition, nutrition analysis, portion size detection and the extraction of health-related statistics.

Food category recognition from images poses significant challenges due to the inter- and intra-class complexities inherent in food images. Recent advancements in image recognition marked the emergence of general frameworks like YOLO [10] architectures with high accuracies. However, generalized frameworks encounter difficulties when applied to specific data domains, where inter-class similarities are more pronounced, as in images of dishes prepared with similar ingredients. Additionally, food images exhibit high intra-class variability in terms of visual appearance due to factors like camera angle, portion size, the arrangement of side ingredients, and the visual presentation of the dish. Among these challenges, most state-of-the-art recognition frameworks achieve overall accuracies around 60%-78% on food datasets such as Food101, UEC256. Meanwhile, these frameworks often show poor performance for certain food categories, with recognition rates as low as 20%-40%. As a result, achieving top-1 recognition accuracy remains a demanding goal, and most research focuses on top-5 recognition results, where uncertainty still exists.

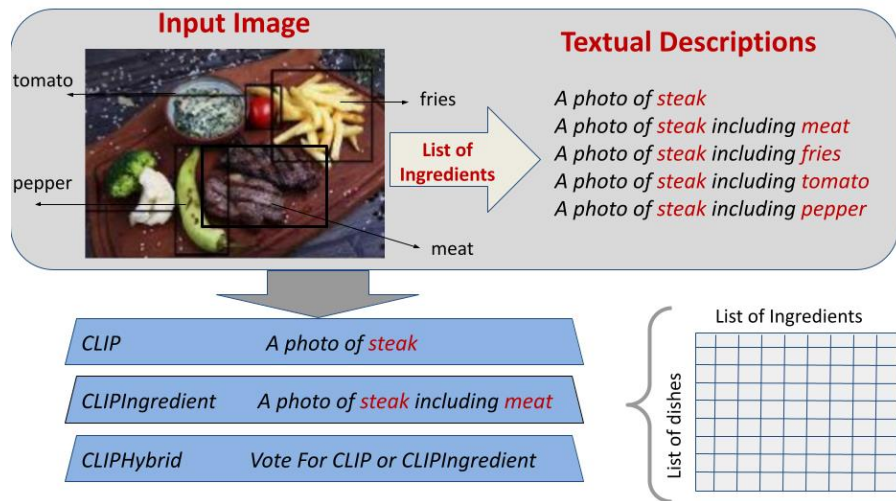
### 3. Food Image Recognition Framework

To overcome the variations in visual appearance within the same categories and enhance the discrimination power of recognition under high similarity conditions, we propose a fusion approach that combines both image and textual cues from single food images by using the CLIP [13]. Key contributions of this work are twofold: 1) Using small window patches for classification of food material rather than food type to model ingredient description. 2) Employing CLIP (Fig. 2) architecture for fusion of food image and ingredient text.

**Fig. 1** presents the proposed architecture, where the first step involves extracting ingredients from recipe data and associating it to food categories. Associated recipes are processed to find the significant key words representing ingredients. Then a voting mechanism is then applied to select the most critical ingredients from the detected ones. Finally, these selected ingredients are used to generate textual descriptions, which are then fused with the image data in the CLIP model to enhance recognition performance. Fig.2 provides an overview of the CLIP architecture, where text and image information are processed in separate encoders and then jointly aligned together through a correlation step to feed the final output for recognizing the content. This architecture helps the model focus on the most relevant ingredients, improving the quality of the image-text fusion. The voting mechanism ensures that the most important ingredients are prioritized, reducing noise from less significant ones. Ingredient detection is an interesting and challenging problem in computer vision, presenting domain-specific difficulties. Unlike other object-part recognition tasks, ingredients can vary significantly in visual appearance, leading to diverse presentations of the same ingredient. Researchers in the field of ingredient detection focus on identifying all ingredients with clear boundaries within a dish. In this work, the primary goal is to detect the most prominent ingredients, which can serve as cues for further analysis. This helps us to obtain an efficient mechanism, where noise is highly regarded and purpose-guided context is extracted from images. In another words, only most reliable one or two ingredients are passed to the next fusion step.

### 3.1. Fusion of Image and Ingredients Data

As illustrated in **Fig. 3**, a custom ingredient labeling effort was conducted to construct mid-sized ingredient regions using the YOLO Annotation tool. These labeled regions are used for both training and performance testing of the YOLO architecture on a set of 101 classes from the Food101 dataset [14], which has been explored in other work of ours. This set of ingredients differs slightly from the standard ingredient names used in [21]. For example, meat dishes, sauce types, and vegetables are organized in greater detail, and some ingredients are assigned different names. A part of ingredient list can be given as: tomatoes, bread, pork, beef, apple, butter, chocolate, sauce, onion, meat, greenery, rice, chicken, fish, potato, soup, dough, egg, sausage, cake, cheese, spaghetti, cake, brown rice, crab, noodles, chili, sausage, bacon, ...



**Figure 3.** Fusion of Image and Textual Descriptions.

“Information fusion of image and text is carried out in three distinct ways when feeding data into the CLIP architecture. In the first approach, the standard CLIP implementation is used, which relies on class-name-based retrieval. In the second approach, a textual description is generated along with the class name and the dominant ingredient identified in the first step. The third approach employs a confidence voting algorithm to select between the first or second approach. Confidence values are generated as binary votes (0 or 1), determining whether the first or second approach should be used, based on a set of training and testing evaluations. At this point, the dataset characteristics influence the decision-making system. Depending on the specific photos in the dataset, class names alone may provide better results, or a combination of ingredients and class names may yield better performance. **Fig. 3** demonstrates the sample textual descriptions which are scanned over a grid of food category vs ingredients. Since ResNet50 and ViT-B/32 are the two common architectures within CLIP that are used in food recognition domain, we also utilize these two architectures for our system. Implementation details and experimental results are given in the next section proving the critical role of the proposed system in addressed challenges. **Table 1** gives average recognition performances of the proposed architecture.

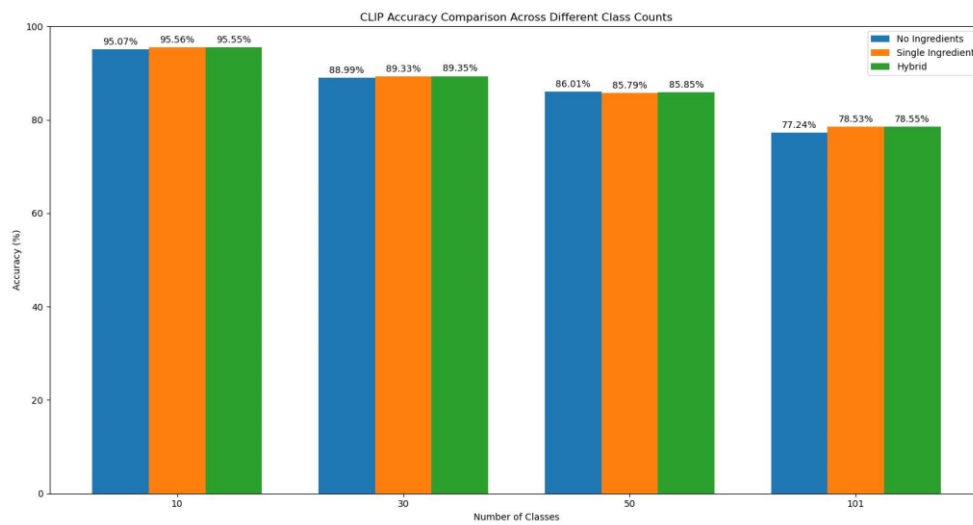
**Table 1.** Results of Food101 with 25000 images in total.

| Method                      | Accuracy (%) |
|-----------------------------|--------------|
| Yolo v8                     | 81.46        |
| CLIP-ResNet50 w class names | 77.27        |
| CLIP-ResNet50 w Ingredient  | 78.44        |
| CLIP-ResNet50 w Hybrid      | 78.38        |
| CLIP-ViT-B/32 w class names | 80.71        |
| CLIP-ViT-B/32 w Ingredient  | 81.79        |
| CLIP-ViT-B/32 w Hybrid      | 81.81        |

## 4. Experimental Results

To evaluate the feasibility of the proposed method, extensive experiments were conducted on the Food101 dataset to analyze the effectiveness of textual ingredient fusion on both the overall recognition accuracy and

class-based improvement rates. The Food101 dataset, which consists of 101 types of dishes with a total of 101,000 images, is used as the primary dataset throughout all experiments. Experiments are organized into four sub-groups based on the number of classes: 10, 30, 50, and 101. The average accuracy per sub-group is presented in **Fig. 4** for three different approaches: the CLIP baseline, CLIP with ingredients, and the hybrid approach. Since fewer classes have a limited capacity to represent the variety of food types, only a small increase in accuracy is observed overall. The CLIP baseline achieves 95.07% accuracy across 10 classes, while it achieves 88.99% accuracy across 30 classes. When ingredient fusion with images is applied, the overall accuracy increases to 95.55% for 10 classes and 89.35% for 30 classes. With 101 classes considered for inspection, the improvement in overall accuracy becomes more apparent, as the recognition accuracy increases from 77.24% to 78.55% when the ResNet50 architecture is utilized. To observe the performance improvement of the proposed architecture across different dish types, the top 12 dishes showing the highest improvement rates are displayed in **Table 2**. The learning rate depends on the number of available images. However, an excessive amount of data can increase the tendency toward biased results. To analyze how the number of images per category affects the final recognition results of the proposed system, 100, 250, 500, and 1000 images per class were distributed across four experimental setups. All four groups showed improvements in accuracy, with 250 images per class setup achieving the highest overall increase. This suggests that a simpler and faster setup is optimal for evaluation and practical use with the Food101 dataset.



**Figure 4.** Class numbers-based performance comparison.

**Table 2.** Performance Comparison of SOTA Methods on Food101 dataset.

| Increase Rates  |       | Method                                 | Accuracy (%) |
|-----------------|-------|----------------------------------------|--------------|
| Beet Salad      | 26.8% | Yolo v8 [25]                           | 81.46        |
| Pork Chop       | 22.4% | CNN-HC-FT [17]                         | 79.78        |
| Strawberry Cake | 14%   | Foster w Comp [13]                     | 79.56        |
| Hamburger       | 10.4% | MVFSL-TC [12]                          | 55.3         |
| Lasagna         | 9.2%  | DeepFood [8]                           | 77.4         |
| Eggs benedict   | 6.8%  | Fusion Learning [29]                   | 68.76        |
| Pizza           | 4.8%  | CLIP-ViT-B/32 w Hybrid ( <b>Ours</b> ) | 81.81        |

#### 4.1. Implementation Details

We employ zero-shot classification implementation of CLIP based on two pre-trained models: ResNet50 and Vision Transformer (Vit-B/32) with batch size of 32 and a varying number of classes as 10,30,50,101. ResNet-50 takes approximately 0.5secs per batch, whereas Vit-B/32 takes 0.7secs on a computer with an Intel i7 processor and graphics card RTX3090. Processing full images takes around 28 minutes for ResNet50 and 36 minutes for Vit-B/32 for each evaluation cycle. For data processing, standart CLIP pipeline is followed and images are resized to 224x224. PIL (Python Imaging Library), CLIP's built-in processing functions and PyTorch dataset and data loader are used for supporting libraries. Table 2 depicts the performance improvement



compared to the CLIP baseline and hybrid approach. Ingredient fusion and the hybrid approach yield similar results most of the time, demonstrating the positive contribution of ingredient cues to the results. As shown in the figure, the “beet salad” class benefits the most from the ingredient context, with its accuracy increasing from 50% to 78%. The second largest improvement is observed in the “pork chop” class, where accuracy increases from 30% to 55%. The “eggs benedict” and “pizza” classes show a smaller increase when textual ingredient descriptions are included. This could be due to the nature of the ingredients detected in these dishes, which may improve results for certain food images but have little to no effect on others.

## 5. Conclusions

In this work, we propose enhancing recognition performance by extracting contextual information from images within the food domain. Our approach incorporates a CLIP based image and text fusion architecture where the first step involves extracting characteristic ingredients from recipes. The final performance of the framework is evaluated through extensive experiments by utilizing ResNet50 and ViT-B/32 372 image encoders within CLIP on Food101 dataset. Experimental results demonstrate the effectiveness and efficiency of the proposed architecture in significantly boosting recognition performance, with improvements ranging from 0.5% to 15%. Our approach notably outperforms existing techniques, surpassing even YOLO v8 in overall food category recognition with an overall accuracy of 81.81%. This work highlights a pioneering contribution by leveraging contextual information for food image understanding, with the potential for application in various domain-specific image recognition challenges. It also presents significant promise in advancing fields such as nutrition, culinary science, food science, and health.

## Declaration of interest

The authors declare that there is no conflict of interest.

## Acknowledgements

The author gratefully acknowledges the support provided by Ela Sava and Münib Akar on analyzing images.

## References

- [1] Chen X, Kamavuako EN. “Vision-based methods for food and fluid intake monitoring: A literature review”, *Sensors*, (2023) 23(13), 2023.
- [2] Ponte D et al. “Ontologydriven deep learning model for multitask visual food analysis”, *VISIGRAPP* (2024) 624-631.
- [3] Zhang Y et al. “Deep learning in food category recognition”, *Information Fusion*, (2023) 98:101859.
- [4] Zhao H et al. “Fusion learning using semantics and graph convolutional network for visual food recognition”, In *WACV*, (2021) 1710–1719.
- [5] Liu C et al. “Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment”, In *International Conference on Smart Homes and Health Telematics*, Springer Intl. Publishing, (2016) 37–48.
- [6] Shuqiang J et al. “Few-shot food recognition via multi-view representation learning”, *ACM Transactions on Multimedia Computing, Communications and Applications*, (2020). 1-4.
- [7] Yang J et al. “Learning to classify new foods incrementally via compressed exemplars”, *CVPRW*, (2024) 3695-3704.
- [8] Ergun OO, Ozturk B. “An ontology based semantic representation for turkish cuisine”, In *26th Signal Processing and Communications Applications Conference (SIU)*, (2018) 1–4.
- [9] Morales R et al. “Robust deep neural network for learning in noisy multilabel food images”, *Sensors*, (2024) 24(7).
- [10] Jocher I et al. *Yolov8*, 2023.
- [11] Mao R et al. “Visual aware hierarchy-based food recognition”, In *ICPR*, (2021) 571–598.
- [12] Morales R, Quispe J, Aguilar E. “Exploring multi-food detection using deep learning-based algorithms”, In *IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, (2023) 1–7.
- [13] Radford A et al. “Learning transferable visual models from natural language supervision”, *arxiv*, (2021). Available: <https://arxiv.org/abs/2103.00020> (accessed: May 5, 2025).
- [14] De-Vera R et al. “Lofi: Long-tailed fine-grained net- 433 work for food recognition”, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, (2024) 3750–3760.
- [15] Zhang W et al. “Hsnn: A subnetwork-based encoding structure for dimension reduction and food classification via harnessing multi-cnn model high-level features”, *Neurocomputing*, (2020) 57–66.
- [16] Aguilar E, Nagarajan B, Radeva P. “Uncertainty-aware selecting for an ensemble of deep food recognition models”, *Computers in Biology and Medicine*, (2022) 146:105645.
- [17] Alahmari S et al. “Segment anything in food images”, In *CVPRW*, (2024) 3715–3720.
- [18] Bossard L et al. “Food-101 – mining discriminative components with random forests”, In *ECCV*, (2014) 446–461
- [19] Ponte D et al. “Multi-task visual food recognition by integrating an ontology supported with llm”, In *SSRN*, (2024) 3695–3704.
- [20] Che C et al. “Enhancing multimodal understanding with clip-based image-to-text transformation”, In *Proc. of the 6th International Conference on Big Data Technologies*. Association for Computing Machinery, (2023).
- [21] Min W et al. “Ingredient-guided cascaded multi-attention network for food recognition”, In *Proc. of the 27th ACM*

- International Conference on Multimedia, (2019) 1331–1339.
- [22] Chen J, Ngo C. “Deep-based ingredient 391 recognition for cooking recipe retrieval”, In Proceedings of the 24th ACM International Conference on Multimedia, (2016) 32–41.
  - [23] Cozzolino D et al. “Raising the bar of ai-generated image detection with clip”, In CVPRW, (2024).
  - [24] Li M et al. “Clip-event: Connecting text and images with event structures”, 2022.
  - [25] Ganz R, Elad M. “Text-to-image generation via energy-based clip”, 2024.
  - [26] Zhang Z et al. “Dual-image enhanced clip for zero-shot anomaly detection”, 2024.
  - [27] Sain A et al. “Clip for all things zero-shot sketch based image retrieval, fine-grained or”, CVPR, (2023)2765–775.
  - [28] Rawlekar S “et al. Prior-aware multilabel food recognition using graph convolutional networks”, In Extended Abstract in MetaFood, CVPRW, (2024) 3695–3704.
  - [29] Wu Y et al. “Few-shot food recognition with pre-trained model”, In Proc. of the 1st Intl. Workshop on Multimedia for Cooking, Eating, and Related APplications”, ACM (2022) 45–48.