Khana: A Comprehensive Indian Cuisine Dataset

Omkar Prabhu
prabhuomkar@pm.me

Abstract

As global interest in diverse culinary experiences grows, food image models are essential for improving food-related applications by enabling accurate food recognition, recipe suggestions, dietary tracking, and automated meal planning. Despite the abundance of food datasets, a noticeable gap remains in capturing the nuances of Indian cuisine due to its vast regional diversity, complex preparations, and the lack of comprehensive labeled datasets that cover its full breadth. Through this exploration, we uncover `Khana`, a new benchmark dataset for food image classification, segmentation, and retrieval of dishes from Indian cuisine. Khana fills the gap by establishing a taxonomy of Indian cuisine and offering around `131K` images in the dataset spread across `80` labels, each with a resolution of `500x500` pixels. This paper describes the dataset creation process and evaluates state-of-the-art models on classification, segmentation, and retrieval as baselines. Khana bridges the gap between research and development by providing a comprehensive and challenging benchmark for researchers while also serving as a valuable resource for developers creating real-world applications that leverage the rich tapestry of Indian cuisine.

Webpage: https://khana.omkar.xyz

1 Introduction

In this digital age, food transcends the physical realm due to the exponential rise of smartphones and social media with vibrant pictures. With this rise, there is a need for efficient navigation across diverse culinary landscapes for several delivery and food review platforms, as well as for precise dietary assessment and personalized nutrition. Food image classification and food retrieval plays a critical role here, but the journey from pixel to palate presents unique challenges for all cuisines. [11, 19, 20] impacted food classification by establishing benchmark datasets, pioneering architectural advancements, and showcasing practical implementations. They have significantly enriched the field by introducing pivotal datasets and innovative architectures while highlighting various challenges encountered in fine-grained categorization throughout the years.

Indian cuisine, in particular, is a kaleidoscope of flavors and textures. Different spices, textures, and ways of cooking create several delicious varieties of food. The close resemblance often masks these unique flavors and details. This heterogeneity becomes a significant hurdle for image classification algorithms trained on broader datasets. The need for detailed categorization in a convoluted domain like food emphasizes the importance of fine-grained image classification, which involves recognizing subtle visual nuances. Food classification has seen considerable effort, especially in Western and Asian cuisines. Even in the case of Asian cuisines, the emphasis has predominantly been on Japanese or Chinese influences. The research conducted thus far has significantly contributed to the advancement of food classification within these specific culinary domains.

Figure 1: Representative images from each category

Recognizing this gap, we introduce Khana, a comprehensive benchmark dataset for fine-grained food image classification in the Indian context, as the main contribution. It contains `131K+` images with `80` categories belonging to different super-classes, such as breakfast, main course, snacks, and beverages. Beyond its sheer size and curated composition, Khana paves the way for a future where the nuances of Indian food can be readily interpreted by machines, bridging the gap between the visual and the delectable. In addition, we provide taxonomy with extensive experiments comparing various state-of-the-art methods of image classification. We hope to empower research, fuel innovation, and celebrate the diversity and richness of Indian food, one pixel at a time.

2 Related Work

This section highlights prior noteworthy research focusing on food-related datasets, work on food classification, segmentation and retrieval.

2.1 Food Related Datasets

Food datasets as shown in Table 1 have grown in recent years, covering different cuisines, cooking styles, and annotation formats developed for food-related tasks like classification, segmentation, and retrieval. Most of these datasets consist of Eastern and Western food dishes, with very little (mostly fine-grained) work done on cuisines like Chinese [4], Japanese [21], Brazilian [7], Kenyan [10], and Singapore [24]. Despite the global prevalence and increasing popularity of Indian cuisine [2], it remains underrepresented in food-related research work. The existing research lacks representation of Western dishes adapted to Indian tastes, distinctive regional cooking

Dataset	Images	Categories	Cuisine	Year
ChineseFoodNet	180K	208	Chinese	2017
THFOOD-50	15K	50	Thai	2017
Food524DB	247K	524	Misc.	2017
Food-101N	310K	101	Misc.	2018
KenyanFood13	8K	13	Kenyan	2019
FoodX-251	158K	251	Misc.	2019
SUEC Food	32K	256	Asian	2019
Sushi-50	3.9K	50	Japanese	2019
FoodAI-756	400K	756	Singapore	2019
ISIA Food-500	399K	500	Misc.	2020
MyFood	1250	9	Brazilian	2020
FoodSeg154	10K	154	Misc.	2021
Food2K	1M	2000	Misc.	2023
DailyFood-172	42K	172	Misc.	2024
AI4Food-NutritionDB	558K	893	Misc.	2024
Khana	**130K**	**80**	**Indian**	**2025**

Table 1: Comparison of the Khana dataset with prior related works

techniques, and the vibrant cultural elements that characterize Indian cuisine. Although the datasets generated until now are either scraped from the web or captured using manual labor, they do not tap into the widespread network of social media and food delivery applications. While the existing research is valuable, it lacks coverage of Indian cuisine, creating a need for exploratory analysis and benchmarking that would benefit the research community.

2.2 Food Classification and Segmentation

Food classification and segmentation are becoming more prominent in recent years, with real-world applications like dietary assessment, automated food logging, and nutritional analysis. Foundational work from [3] established a benchmark with 101 food categories that has driven later research in automated food recognition. There were early deep learning experiments [13] that demonstrated the benefits of CNNs for dietary assessments. We see a gradual evolution of fine-grained classification with regional cuisines like [4], [11], and transfer learning like [12]. We also see large-scale neural networks for food recognition in [19], [20], and [15]. For faster latency mobile use cases, there was a novel idea in [28]. The transition from classification to segmentation happened via [31], which enabled pixel-level food understanding. Contemporary research for real-world deployment challenges was presented in [15], and nutrition-focused recognition systems with [23] and [22]. Along with all these, [14] highlights the transition from traditional computer vision to sophisticated deep learning architectures for food classification and segmentation tasks.

2.3 Food Retrieval

Food retrieval has become more prevalent in recent years due to its useful applications, such as dietary management, recipe recommendation, restaurant services, and food logging. [30] uses bi-directional LSTMS for encoding recipes and images in a common embedding space, but it struggles with noisy images and cross-domain matching between images and texts. [5] uses CNNs for food retrieval based on embeddings with distance metrics, and it points out challenges due to noisy and cluttered food images, large intra-class variations due to differences in preparation, portion size, and presentation, as well as high inter-class similarity among visually overlapping dishes. [18] proposes bi-directional retrieval using CNNs for visual and deep language models to process ingredients and instructions. It suffers from a gap between visual and textual representations, variability in food presentation and image quality, and incomplete or ambiguous recipe descriptions. [29] proposes adversarial training for CNNs to learn cross-modal embeddings that map food images with recipe text for retrieval. It addresses challenges like domain shift between visual and textual modalities, noisy and ambiguous recipe descriptions, and the high visual similarity of different dishes. [25] focuses on using CNNs to extract visual features and learning a similarity metric to improve retrieval performance. [19], designed to advance recognition and retrieval tasks, uses global attention to capture holistic dish appearance and local attention to highlight discriminative regions such as textures or ingredients, which helps in large-scale settings. Most of the studies highlight challenges such as large intra-class variation due to different preparation and presentation styles, high inter-class similarity among visually related dishes, and noisy image conditions that make retrieval difficult.

3 Khana Dataset

Given the considerable progress in food domain, existing datasets exhibit a clear lack of diversity. In particular, Indian cuisine, that is one of the largest and most varied culinary traditions remains underrepresented. To overcome this limitation, we present the research community with a novel single modality self-collected dataset and arranged using smart taxonomoy.

3.1 Establishing a Taxonomy for Indian Cuisine

The purpose of the taxonomy is two-fold: First, it aims to organize and structure the food items by establishing hierarchical relationships and culinary categories based on their preparation methods, regional origins, and cultural significance. Second, the taxonomy provides well-defined categories and subcategories that serve as training labels for usability in tasks like classification, segmentation, and retrieval. The taxonomy facilitates semantic search capabilities by establishing relationships between similar food items and cooking techniques.

Figure 2 shows the hierarchy and the number of images per dish variety. Each label is a dish variety in the dataset that belongs to a food category i.e. `category` → `dish` → `variety`. Dishes are categorized based on their ingredients, cooking methods, and regional cuisine. For example, *dosa*, *idli*, *uttapam*, *medu vada* are all dishes made up of ingredients like lentils and rice categorized as *south indian* based on their regional nature. For example, *anda curry*, *chana masala*, *fish curry*, *bhindi masala* are all dishes categorized as *curry* based on their cooking method and food consistency. Pav and chutney emerge as integral components, contributing to the dataset's diversity. *pav*, a type of bread roll, accompanies several food dishes across India in different forms. The dataset unfolds a fascinating narrative of how one dish can manifest based on its geographical origin. *aloo*-related, *puri*-related, and *bread*-centered dishes exhibit distinct regional adaptations, reflecting the cultural nuances and local preferences as famous in different

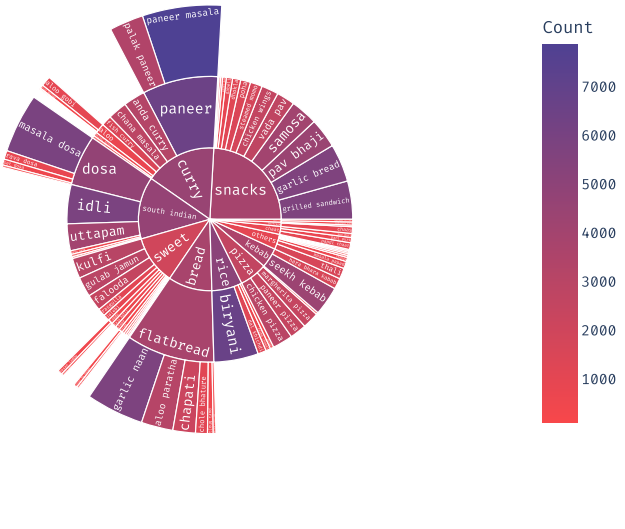


Figure 2: Hierarchical sunburst showing categories, dishes, varieties, and image counts per dish variety

parts of India. For instance, *aloo*-related dishes may vary in preparation and spice profiles, with regional influences shaping the culinary identity of each sample.

The taxonomy is built with flexibility and scalability at its core, where it can expand as culinary traditions evolve or as specific focus areas gain prominence. The categories and dishes can grow without disrupting the existing hierarchy. For example, *curry* category can be expanded to different dishes based on different regions. Food item innovations in urban areas can be expanded into *snacks* category. It also addresses cross-cultural and fusion dishes through flexible categorization, such as *pizza* with Indian adaptations *paneer pizza*. It also has traditional varieties and dishes that span different categories like *steamed momo* and *seekh kebab* tagged as both veg and non-veg under dietary preferences to reflect their nature.

3.2 Data Collection and Cataloguing Addressing Multilingual Conventions

The dataset showcases rich culinary diversity, spanning appetizers, main courses, desserts, snacks, and beverages. It features iconic dishes like *dosa*, *biryani*, *gulab jamun*, and *chaas*, with a wide variety of dishes such as *flatbreads*, *curries*, and *sweets*. Both vegetarian and non-vegetarian options, including *paneer* dishes and *fish curry*, are well-represented, highlighting regional flavors and dietary preferences. Special categories like *fasting foods* and *kebabs* further enrich the taxonomy, reflecting the breadth of Indian and global cuisine.

The data for the Khana dataset comes from search engines and online food delivery platforms like Swiggy and Zomato [1]. Web crawlers gathered images in an automated manner from keyword search results and restaurant delivery menu lists. The process removed duplicate images across restaurant menus and search results for varieties of the same dish by finding nearest neighbors using image embeddings generated from torchvision models [17]. Simple scripts filtered out low-quality images that did not meet the resolution criteria.

The images in the dataset are structured as per the dish variety names and there is a taxonomy CSV which contains information like `category`, `dish`, `variety`, `dietary` for each label in the dataset. There is no additional information such as preparation method or timestamps available.

We labeled the images using an automated approach during the collection stage, based on keyword searches or substring matches with restaurant menu item names. To ensure consistent labeling, we grouped samples containing varied Hinglish keywords for the same dish variety. For example, the label *pani puri* can be also denoted as *pani poori*, *golgappa*, *panipuri* or *panipoori*. We filtered out images that contained a combination of different labels. To ensure label accuracy, the image folder for each label was manually verified by three annotators, who achieved inter-annotator agreement for classifying certain samples. For example, lot of *south indian* dishes were available as a single combo-item in the dataset, which needed to be filtering.

There were no image processing techniques such as rotation, flipping or color adjustments used on the dataset to ensure its originality. There are no augmented samples in the dataset.

3.3 Dataset Statistics and Characteristics

The dataset contains around `131K` images spread across `80` different classes. Each sample in the datset has a resolution of `500x500` pixels. The dataset is split into training, validation, and test sets with `70%` train, `15%` validation, `15%` test respectively. The dataset exhibits an imbalanced class distribution, with some food categories and dishes having a higher number of samples, while others are underrepresented. For example, popular dishes like *masala dosa* and *biryani* have more images, whereas dishes like *neer dosa* and *chikki* have fewer samples. The average number of samples per class varies, and this imbalance requires techniques like data augmentation for better model performance. Figure 3 shows the number of images per food dish from the dataset.

Figure 5 showcases the diverse visual representations within a single food dish. It is crucial to include these variations when creating datasets for food recognition as they enable more accurate and generalizable models. Figure 4 illustrates striking visual similarities between two distinct food dishes. These examples aid in training models to achieve precise classification accuracy and make them resilient to misclassification errors.

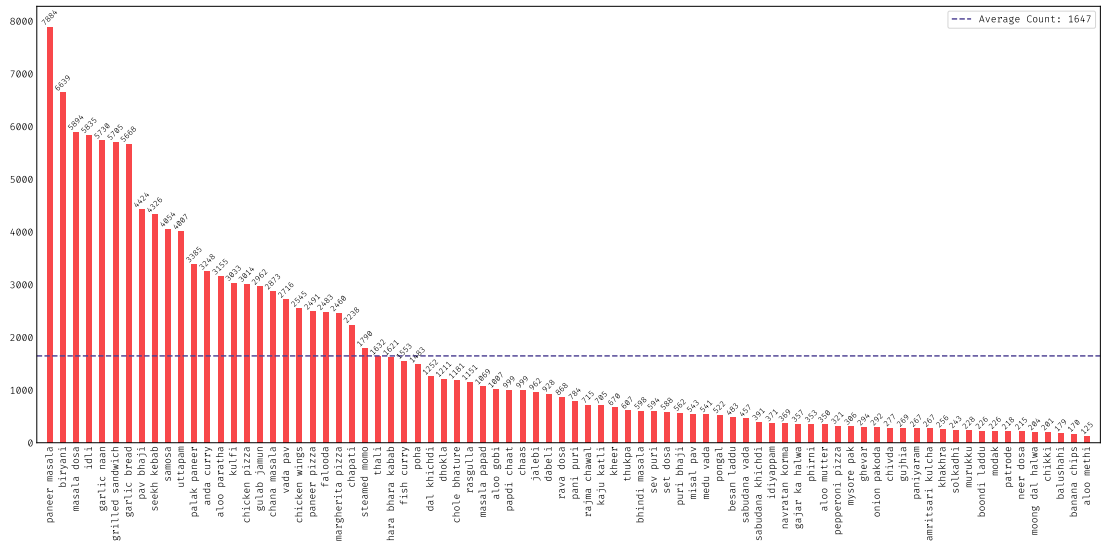


Figure 3: The distributions over each category

(a) Similarities *misal pav* and *pav bhaji*

(b) Similarities of *vada pav* and *dabeli*

Figure 4: Similarities between food dishes

Figure 5: Different visual representations of single food dish

4 Experiments

4.1 Experimental Setup

In this section, we outline the image classification baselines selected for comparative analysis and the ideas driving their selection along with the setup for running the experimental analysis. Our emphasis was predominantly on leveraging pre-trained architectures like Convolutional Neural Networks (CNNs) and Transformers that are known for their proficiency in image classification problems. We opted for following models:

Model	Total Params	Trainable Params
ResNet-152	58,246,258	102,450
EfficientNet-V2-S	20,241,538	64,050
ViT-B-16	85,837,106	38,450
ConvNeXT-S	49,493,138	38,450

Table 2: Comparison of image classification model parameters

- **Residual Networks (ResNet)**: ResNet [9] have proven to be a strong baseline for comparison in computer vision research. We used this deep convolutional neural network architecture known for its residual learning approach and leveraged the pre-trained weights of *ResNet-152*, which capture essential generic image features from the extensive ImageNet dataset and provide a robust starting point. It facilitates more efficient learning of task specific features with an unbalanced food dataset. Additionally, it might not be an optimal choice for all classification tasks due to potential overfitting.

- **EfficientNet**: EfficientNetV2 [26, 27] is a lightweight and efficient convolutional neural network architecture known for achieving high accuracy while requiring fewer parameters and computational resources when compared to other models. It utilizes a compound scaling method, uniformly scaling network depth, width, and resolution to maintain efficiency. EfficientNet has proved to be an industry accepted solution for small to medium scale image classification tasks. We used *EfficientNet-V2-S* variant pre-trained on ImageNet for learning specific features from our dataset quickly, serving as a strong starting point for further improvement.

- **Vision Transformer (ViT)**: Vision Transformers [6] has demonstrated high accuracy on various benchmarks, making it a valuable benchmark for comparison. It doesn't rely on pre-defined assumptions about image features, potentially leading to better generalization, and requires fewer resources when compared to CNNs of similar accuracy. We chose *ViT-B-16* or *ViT-Base* variant pre-trained on ImageNet as a competitive baseline for comparison, allowing us to benchmark and validate its effectiveness.

- **ConvNeXT**: ConvNext [16] utilizes standard convolutional modules without relying on self-attention mechanisms as in transformers, leading to a simple and more interpretable architecture. It has achieved competitive results on various image classification tasks, proving to be a strong baseline for comparison. We leveraged *ConvNeXT-Small* or *ConvNeXT-S* pre-trained on ImageNet as a baseline as it contains comparable model parameters. ConvNeXT offered a modular design, depthwise separable convolutions, and a residual-like set aggregation block, which gave a balance between accuracy and efficiency. These chosen baselines offer a diverse set of architectural choices and complexity levels, providing a comprehensive comparison for evaluating the performance of our proposed approach on the food dataset.

These chosen baselines offer a diverse set of architectural choices and complexity levels, providing a comprehensive comparison for evaluating the performance of our proposed approach on the food dataset. The experiment encompasses of fine-tuning a pre-trained model, which involves selectively immobilizing the weights of earlier layers while focusing training efforts solely on the final layers customized for the specific classification task. Optimization of the model is facilitated through the `Adam` optimizer, leveraging a learning rate parameter set at `0.001`. The loss

function utilized is cross-entropy, a widely acknowledged metric in classification tasks. Training unfolds over a duration of `50` epochs, with a batch size of `64`, to iteratively refine the model's performance. Our setup defines a transformation for pre-processing images. This transformation resizes images and then extracts a central square matching the model's expected size as shown in 3 It ensures consistency and avoids exceeding model limitations. Pixel values are normalized using values from the ImageNet dataset. It involves subtracting the mean and dividing by the standard deviation for each color channel (`red, green, blue`). This process effectively removes common variations in pixel intensity across images, leading to improved model performance and convergence. Finally, for smooth rescaling during the resizing process we use bilinear interpolation.

4.2 Results

Out experimental analysis included comparing four state-of-the-art (SOTA) models: *ResNet, EfficientNet, ViT, and ConvNeXT*. We evaluate the performance of each model based on loss curves and top-1 and top-5 accuracy. Figure 6 shows the loss curves for the proposed method and each SOTA model during training. All models achieved convergence within `50` epochs. ViT reached the lowest final loss of `0.2`, followed by ResNet with a loss of `0.3` and ConvNeXT with a loss of `0.4`, while EfficientNet showed a slightly slower and less stable convergence behavior, reaching a final loss of `0.7`. Table 3 presents the top-1 and top-5 accuracy results for all models on the Khana dataset. *ConvNeXT-S* model achieved the highest top-1 accuracy (`86.72%`) and top-5 accuracy (`97.58%`), outperforming the SOTA models with the lowest margin of `1.4%` and `0.4%`, for top-1 and top-5 accuracy, respectively.

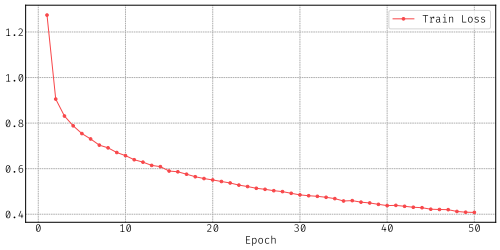
Model	Crop Size/Resize Size	Top-1 Accuracy	Top-5 Accuracy
ResNet-152	224/232	81.00	95.37
EfficientNet-V2-S	384/384	80.47	95.52
ViT-B-16	224/256	85.34	97.15
ConvNeXT-S	**224/230**	**86.72**	**97.58**

Table 3: Comparison of top-1 and top-5 accuracy for baselines

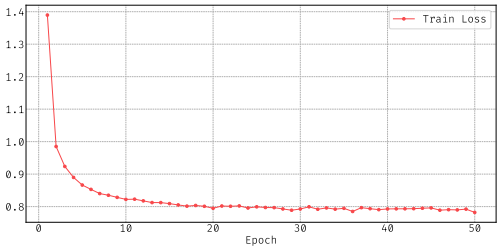
5 Limitations

The dataset suffers from class imbalance, as evident from Figure 3, with several popular food categories overrepresented and very niche food items underrepresented that leads to potential bias. Standard evaluation metrics do not capture fine-grained distinctions between food categories, and preprocessing can further influence the results. The dataset is promising for our given scope, but using it in applications to other contexts or datasets with different cultural and environmental conditions may lead to variability in results.

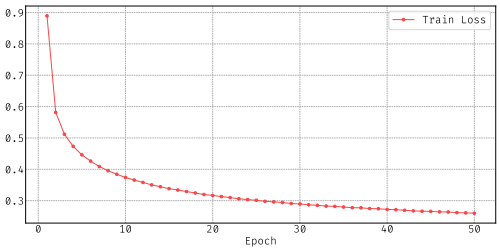
Exploring beyond this study's scope, we identify that adding more images for underrepresented food categories, expanding to new cuisines, and improving annotations are crucial for making models trained on this dataset more accurate for real-world applications. Given that the existing dataset covers a wide range of categories, we can expand beyond the baselines with new image classification models. Given the rapid growth in usage of multi-modal LLMs, researchers can explore querying over images and conducting qualitative comparisons of embeddings for several multi-modal LLMs as well.



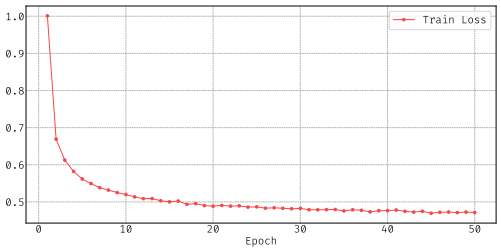
(a) ResNet-152



(b) EfficientNet-V2-S



(c) ViT-B-16



(d) ConvNeXT-S

Figure 6: Training loss curves for baseline models

Acknowledgments

We want to thank Arpita Rane, Akshay Pithadiya, John D'souza, and Jane D'souza for their help in annotating a small sample of images during the development of an initial version of this dataset.

References

[1] Food delivery market size to cross Rs 2 lakh crore by 2030: Bain-Swiggy report — economictimes.indiatimes.com. https://economictimes.indiatimes.com/tech/startups/online-food-delivery-market-to-grow-18-on-year-to-rs-2-lakh-crore-by-2030-bain-report/articleshow/111452013.cms. [Accessed 16-01-2025].

[2] Taste Atlas. These are the 100 Best Cuisines in 2025 - TasteAtlas Awards 24/25 — tasteatlas.com. https://www.tasteatlas.com/best/cuisines, 2024. [Accessed 15-01-2025].

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pages 446–461. Springer International Publishing, 2014.

[4] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. Chinesefoodnet: A large-scale image dataset for chinese food recognition, 2017.

[5] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Learning cnn-based features for retrieval of food images. In *New Trends in Image Analysis and Processing – ICIAP 2017*, page 426–434. Springer International Publishing, 2017.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[7] Charles Freitas, Filipe Cordeiro, and Valmir Macario. Myfood dataset, 2020.

[8] Junyi Gao, Weihao Tan, Liantao Ma, Yasha Wang, and Wen Tang. Musefood: Multi-sensor-based food volume estimation on smartphones, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] Mona Jalal, Kaihong Wang, Sankara Jefferson, Yi Zheng, Elaine O. Nsoesie, and Margrit Betke. Scraping social media photos posted in kenya and elsewhere to detect and analyze food types. In *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*, MM '19, page 50–59. ACM, October 2019.

[11] Parneet Kaur, Karan Sikka, Weijun Wang, Serge Belongie, and Ajay Divakaran. Foodx-251: A dataset for fine-grained food classification, 2019.

[12] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 5447–5456. IEEE, June 2018.

[13] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment, 2016.

[14] Detianjun Liu, Enguang Zuo, Dingding Wang, Liang He, Liujing Dong, and Xinyao Lu. Deep learning in food image recognition: A comprehensive review. *Applied Sciences*, 15 (14), 2025. URL https://www.mdpi.com/2076-3417/15/14/7626.

[15] Guoshan Liu, Yang Jiao, Jingjing Chen, Bin Zhu, and Yu-Gang Jiang. From canteen food to daily meals: Generalizing food recognition to more practical scenarios. *IEEE Transactions on Multimedia*, page 1–10, 2024.

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. URL https://arxiv.org/abs/2201.03545.

[17] TorchVision maintainers and contributors. TorchVision: PyTorch's Computer Vision library. https://github.com/pytorch/vision. [Accessed 16-01-2025].

[18] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images, 2018.

[19] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20. ACM, October 2020.

[20] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9932–9949, August 2023.

[21] Jianing Qiu, Frank Po Wen Lo, Yingnan Sun, Siyao Wang, and Benny Lo. Mining discriminative food regions for accurate food recognition. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 158. BMVA Press, 2019.

[22] Sergio Romero-Tapiador, Ruben Tolosana, Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, Isabel Espinosa-Salinas, Gala Freixer, Enrique Carrillo de Santa Pau, Ana Ramírez de Molina, and Javier Ortega-Garcia. Leveraging automatic personalised nutrition: food image recognition benchmark and dataset based on nutrition taxonomy. *Multimedia Tools and Applications*, April 2024.

[23] Doyen Sahoo, Wang Hao, Shu Ke, Wu Xiongwei, Hung Le, Palakorn Achananuparp, Ee-Peng Lim, and Steven C. H. Hoi. Foodai: Food image recognition via deep learning for smart food logging. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2260–2268. Association for Computing Machinery, 2019. ISBN 9781450362016.

[24] Doyen Sahoo, Wang Hao, Shu Ke, Wu Xiongwei, Hung Le, Palakorn Achananuparp, Ee-Peng Lim, and Steven C. H. Hoi. Foodai: Food image recognition via deep learning for smart food logging. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, page 2260–2268. ACM, July 2019.

[25] Wataru Shimoda and Keiji Yanai. Learning food image similarity for food image retrieval. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 165–168, 2017.

[26] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

[27] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR, 2021. URL http://proceedings.mlr.press/v139/tan21a.html.

[28] Chakkrit Termritthikun, Paisarn Muneesawang, and Surachet Kanprachar. Nu-innet: Thai food image recognition using convolutional neural networks on smartphone. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(2-6):63–67, June 2017.

[29] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee peng Lim, and Steven C. H. Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images, 2019.

[30] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee peng Lim, and Steven C. H. Hoi. Cross-modal food retrieval: Learning a joint embedding of food images and recipes with semantic consistency and attention mechanism, 2020.

[31] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven C.H. Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 506–515. ACM, October 2021.