Dean&Francis
ISSN 2959-6157

Research and Analysis of Image-based Food Identification Technology in Complex Scenarios

Guanbo Liu

International School, Beijing University of Posts and Telecommunications, Beijing, China
2024213684@bupt.cn

Abstract:

Food image recognition, as a vital branch of fine-grained visual analysis, holds significant promise for revolutionizing health management and the food industry. While deep learning models have achieved remarkable accuracy in controlled settings, their performance often degrades in real-world environments due to complex challenges. This paper presents a comprehensive review of food image recognition under these complex scenarios. This paper systematically analyzes the primary obstacles, including environmental disturbances (e.g., lighting and background variations), perspective and structural changes (e.g., occlusion and viewpoint diversity), intrinsic food variations (e.g., non-rigid deformation and inter-class similarity), and stringent system constraints (e.g., real-time and computational limits). Correspondingly, this paper surveys and discusses representative technical solutions, such as attention mechanisms, multimodal fusion, geometric transformation, and lightweight network architectures, highlighting their strengths and limitations. The review concludes that while existing methods have made substantial progress, critical issues like the accuracy-efficiency trade-off, limited model generalization, and inadequate robustness in dynamic extremes remain unresolved. Future research should prioritize enhancing model adaptability in open environments, integrating semantic reasoning with perception, and developing comprehensive evaluation benchmarks to bridge the gap between laboratory research and practical deployment.

Keywords: Food image recognition; complex scenarios; deep learning.

ISSN 2959-6157

1. Introduction

Food is not only essential for human survival but also plays a critical role in health management, chronic disease prevention, and cultural heritage. However, in modern life, the growing demand for healthy eating often conflicts with inefficient dietary recording and nutritional assessment methods. With the rapid advancement of deep learning and computer vision technologies, such as convolutional neural networks (CNNs) [1], food image recognition has emerged as a promising solution to automate and enhance this process. As a key branch of fine-grained image recognition, it can accurately identify food categories, ingredients, and even estimate nutritional content from images, thereby supporting applications in personalized health management, smart food industry systems, and agricultural automation.

Although deep learning-based image recognition has achieved high accuracy under controlled conditions—such as ideal lighting, plain backgrounds, and frontal angles—it faces significant challenges in real-world deployment. These include complex environmental interference (e.g., lighting and background variations), diverse camera angles and occlusions, intrinsic variations in food appearance (e.g., cutting and mixing states), and stringent system requirements for real-time and lightweight performance. While various methods—such as attention mechanisms [2], multi-task learning [3], and multimodal fusion [4]—have been proposed to address these issues, systematic solutions for complex scenarios remain under exploration. Therefore, the research focus is shifting from achieving high accuracy in constrained settings to enhancing model robustness and practicality in real-world environments.

In light of these challenges, this paper provides a comprehensive review of food image recognition technologies in complex scenarios. Section 1 introduces the research background and core challenges. Section 3 analyzes the performance of existing methods under various complex conditions and reviews representative solutions, such as CNN-based segmentation and multimodal fusion, discussing their strengths, limitations, and applicability. Finally, the conclusion summarizes key findings, highlights unresolved issues, and suggests promising future directions, such as few-shot learning and open-environment adaptation.

2. Technical challenges and representative solutions in complex scenarios

This chapter will provide an in-depth analysis of the specific challenges faced in four complex scenarios and highlight representative solutions in recent years. These studies demonstrate how model robustness can be improved by improving model architecture, introducing new mechanisms, or integrating multi-source information.

2.1 Environmental disturbances

Environmental disturbances is one of the most common challenges in food image recognition, mainly including background clutter, lighting variations, and image quality defects. Existing studies have proposed a variety of targeted methods to improve the robustness of the model under harsh imaging conditions.

2.1.1 Background interference

In cafeterias or takeaway scenarios, food is often mixed with tableware, packaging, tabletops, etc. In response to this problem, the researchers proposed a variety of deep learning-based detection frameworks. For example, Zhao et al. proposed a recognition and localization method based on improved YOLOv3 for apple picking robots in the context of complex orchards [5]. In this study, the detection accuracy in complex backgrounds such as branch and leaf occlusion, fruit density, and bagging was significantly improved by designing a 13-layer lightweight network structure, optimizing the anchor point size, and adopting a dynamic threshold adjustment strategy, with a mAP of 87.71% and an accuracy of 97% on the validation set. These approaches demonstrate the potential of deep learning models to deal with real-world background interference, but there is still room for performance degradation in extremely dense small target scenarios.

2.1.2 Variable lighting

Changes in light intensity, angle, and color temperature can significantly affect the apparent features of an image. To address this problem, researchers are working on developing models that are insensitive to light changes. Zhao et al. proposed an autoregressive-texture model to construct brightness and texture confidence intervals by establishing the statistical threshold of pixel brightness perturbation and combining the local binary mode texture features that are insensitive to light, so as to effectively distinguish the differences caused by lighting changes and foreground targets [6]. The framework can achieve accurate target detection under both slow lighting changes (e.g., daytime passage) and fast lighting changes (e.g., cloud occlusion). This method avoids the prediction of the lighting type and simplifies the system design, but the model is still sensitive to background interference under extreme lighting conditions.

2.1.3 Image Quality Defects

Image Quality Defects: Issues such as low resolution, blurring, or inaccurate focus are common in images cap-

tured by users. Super-resolution reconstruction technology is an effective means to improve image quality. Su et al. reviewed various SR methods [7], among which the method based on generative adversarial network can generate images with high visual perception quality and realistic texture by learning a large number of HR-LR image pairs, which is helpful to restore detailed information and improve recognition accuracy. In addition, the SR method based on the attention mechanism can focus on key areas of the image for reconstruction. However, these methods are still prone to artifacts when reconstructing high-frequency details, and the computational overhead is large, which is challenging in scenarios with high real-time requirements

2.2 Perspectives and structural challenges and their solutions

In food image recognition, the diversity of viewing angles (such as overhead, head-up, oblique shots) and structural changes (such as food placement posture, container shape, and local occlusion) are important challenges that lead to the increase of intra-class differences and the decrease of recognition performance. Changes in perspective introduce geometric deformation of the image, while structural changes can cause local feature loss or confusion, especially in scenes with multi-food stacking, partial occlusion, or non-standard placement.

2.2.1 Response to perspective changes

To solve the problem of image deformation caused by the change of perspective, an effective idea is to restore the standard perspective of the object through image matching and geometric correction. The fast large-viewing image matching algorithm (PORB) based on ORB proposed by Zeng Qinghua et al. is representative in this regard. The algorithm simulates the imaging process of the camera at different angles by constructing a perspective transformation model [8], using ORB features for fast coarse matching, and then estimating the homography matrix through the RANSAC algorithm for precise matching, which significantly improves the matching efficiency and robustness at large viewing angles. Experiments show that the algorithm can maintain more than 80% matching accuracy when the viewing angle changes more than 60°, and the calculation speed is more than 10 times higher than that of the traditional ASIFT algorithm [8]. In food image recognition, similar methods can be used to align food images from different shooting angles to reduce the impact of angle changes on feature consistency.

2.2.2 Handling of structural changes and occlusion

When there is occlusion or structure loss in the image, the performance of traditional recognition methods is easy to degrade due to local feature failure. Li Xiaoxin et al. systematically summarized a variety of robust processing methods from subspace regression to deep learning in "Review of Occluded Face Recognition" [9], and the core idea is to suppress the influence of abnormal regions by structured error coding or occlusion perception features. For example:

Subspace regression methods (such as SRC and CRC) model the occlusion as error terms and reconstruct the original features through sparse coding.

Deep feature methods (such as PCANet and DeepID) extract higher-order features with occlusion invariance through multi-level convolutional networks.

Iterative weight error coding (such as CESR and RSC) dynamically adjusts the feature weights to weaken the contribution of occluded areas.

These methods have been verified to be effective in face recognition and can be transferred to food recognition to deal with complex structural changes such as cutlery occlusion, food stacking, and local cutting [9].

2.3 Target Variation Challenges and Their Solutions

The problem of target variation in food image recognition is mainly manifested in the huge differences in the presentation state, appearance and composition of food. Specifically, it includes: (1) large differences within categories - the same type of food has a diverse appearance due to different cooking methods, containers, cutting degrees, and eating stages; (2) High similarity between categories - different dishes may use similar ingredients or present similar visual characteristics; (3) Non-rigid deformation - fluid, semi-fluid or stirred foods (such as porridge, salad, sauce) lack a fixed shape. These variations put forward extremely high requirements for the characteristic invariance learning and discrimination ability of the model.

2.3.1 Diversity of food presentation states

The appearance of the same type of food is significantly different due to factors such as containers, plating art, cutting and stirring, and partial edible states. For example, a bowl of beef noodles may have different visual shapes depending on how stretched the noodles, how much soup there is, and where the beef is placed, while a bitten sandwich is significantly different from a complete state.

Multi-level feature aggregation network: Liang Huagang et al. proposed a multi-level convolutional feature pyramid model to achieve feature extraction from global to local through a three-level cascade network, combined with the attention area localization module to focus on discriminant areas (such as specific vegetable pieces in

ISSN 2959-6157

salads) [10], and achieved a top-1 accuracy of 91.4% on the Food-101 dataset. The model enhances the robustness of scale changes through the characteristic pyramid structure, and effectively alleviates the morphological variation caused by cutting and stirring.

Coarse to Fine Feature Aggregation Framework (CBDTN): Liang Shuang and Gu Yu's team designed a multi-stage network containing Boundary Awareness Module (BAM) and Deformed ROI Pooling (DRP) [11], extracted multi-level features through the ConvNeXt backbone network, and used the Transformer encoder to capture long-distance dependencies, achieving a recognition accuracy of 85.87% on the Food-2K large dataset. This framework is particularly suitable for dealing with boundary blurring caused by the diversity of containers.

Dual-stream convolutional neural networks: In order to solve the problem of large differences within Chinese cuisine categories, a feature fusion model based on dual-flow bilinear network is proposed to extract complementary features through the dual-path design of different receptive fields, and combined with compact bilinear pooling to reduce the computational complexity, and the top-5 accuracy on the ChineseFoodNet dataset reaches 96.32%. This method has strong adaptability to local deformation, such as the shape change of food after cutting.

2.3.2 Interclass similarity and fine-grained identification

Many dishes use similar ingredients (such as tomato egg noodles and tomato egg soup) and are difficult to distinguish based on local characteristics alone. In addition, dishes with similar appearances in different cuisines (such as wontons and dumplings) may belong to different categories, requiring the model to have the ability to capture fine-grained differences.

Large Language Model-Guided Ontological Fusion (LLMO-MFR): The latest research uses GPT-4 to automatically construct food ontology, establish a multi-layer semantic relationship between "dish-food group" (such as "tomato and egg" and "pasta" and "soup"), and integrate it into the EfficientNet-B4 network for multi-task learning. On the VireoFood-172 dataset, the method improves the accuracy of Top-1 by 4.7%, effectively distinguishing between confusing dishes. The introduction of ontological prior knowledge enhances the model's understanding of semantic associations and reduces misjudgments caused by visual similarities.

Fine-grained feature pyramid and attention mechanism: The multi-level convolutional feature pyramid model automatically focuses on discriminating regions (such as unique sauce textures in salads) through the attention area localization network, avoiding over-reliance on the overall appearance. Similarly, the NetRVLAD module in the CBDTN framework reduces intra-class differences and expands the inter-class distance by aggregating local feature descriptors.

Semantically enhanced confrontational training:Combined with generative adversarial network (GAN) to synthesize difficult samples (such as slightly deformed dumpling images), the robustness of the model against deformation is improved through data enhancement. Some studies further introduce recipe text information as auxiliary supervision to strengthen the model's understanding of the logic of ingredient combination.

2.3.3 Fluid and non-rigid deformation treatment

Fluid foods (such as soups, milkshakes) or mixed foods (such as fried rice, salads) lack stable contours, and their appearance changes drastically with the container, light, and stirring state, and the traditional shape feature extraction method is very easy to fail.

Boundary Awareness and Deformation Convolution: The Boundary Awareness Module (BAM) in the CBDTN framework enhances edge response through spatial attention, while the Deformed ROI Pooling (DRP) adaptively adjusts the shape of the receptive field to better capture the irregular boundaries of fluid food.

Visual Transformer vs. Adaptive Segmentation: Salafis et al. used Swin Transformer to generate a Category Activation Graph (CAM) as a prompt to drive the Segment Anything Model (SAM) to complete zero-shot segmentation [12], reaching 0.54 mIoU on the FoodSeg103 dataset. This weakly supervised method reduces the reliance on pixel-level labeling and is particularly suitable for fluid food region extraction that lacks a fixed shape.

Multimodal feature fusion: For non-rigid foods, combined spectral data (e.g., hyperspectral imaging) with RGB images are combined. For example, the adaptive optimization genetic algorithm identifies beef spoilage areas through spectral features with an accuracy rate of 95.8%. Similar ideas can be extended to the quantitative identification of fluid food concentrations and components.

2.4 System constraints

System constraints are one of the key challenges faced by food image recognition in actual deployment, which are mainly reflected in three aspects: real-time requirements, limited computing resources, and dynamic processing. These constraints require the model to maintain high recognition accuracy and low response latency despite limited computing resources, especially in mobile devices, embedded devices, or video streaming scenarios. The following combines the existing technology to propose solutions for these three types of constraints.

2.4.1 Real-time demand

In scenarios such as self-checkout, catering robots, and smart retail, the system needs to complete the recognition within milliseconds, otherwise it will affect the user experience and system efficiency. In order to meet the real-time requirements, the researchers mainly optimize the model speed through lightweight model design and inference acceleration technology.

MobileNetV1 significantly reduces computational complexity by introducing Depthwise Separable Convolution to convert standard convolutions into deep convolutions and point-by-point convolutions, and at the same time [13], the speed and accuracy are flexibly balanced by the width factor (α) and the resolution factor (β), and high-speed inference is realized on the mobile end.

ShuffleNetV2 proposes to replace indirect FLOPs with direct speed indicators [14], and achieves better real-time performance on the CPU side by keeping input and output channels consistent, reducing memory access costs, and avoiding excessive packet convolution, which is especially suitable for low-latency scenarios.

Lightweight variants of YOLOv3 (e.g., combined with MobileNet backbone) achieve real-time processing while maintaining high recall while achieving real-time processing in food inspection tasks through single-stage detection architecture and multi-scale feature fusion in food inspection tasks, and are suitable for multi-target recognition in video streams [15].

2.4.2 Computing resources are limited

On mobile phones, embedded devices, or edge computing nodes, memory and computing power are limited, and the model needs to be small in size and low power consumption. Model compression and dedicated lightweight architecture are the mainstream solutions.

SqueezeNet adopts the Fire module (Squeeze layer + Expand layer) to compress the number of channels by convolutioning 1×1, greatly reducing the number of parameters (model size is only 0.5MB), and retaining the large activation graph combined with the latency downsampling strategy, which still maintains high classification ability on resource-constrained devices [16].

MobileNetV2 introduces a linear bottleneck structure and a reverse residual module on the basis of V1 to avoid information loss of low-dimensional features, further compress the model volume, and achieve the same accuracy as large models on ImageNet, which is suitable for embedded deployment [17].

Model compression techniques such as pruning [18], quantization, and knowledge distillation can be combined with lightweight models to further reduce the amount of parameters and computation. For example, 8-bit quantization of a pre-trained FoodNet model compresses the model volume to 1/4 of the original without losing precision.

2.4.3 Dynamic processing

For the recognition task of video stream or continuous frames, the system needs to handle the correlation of the time dimension and ensure the consistency and stability between frames. Temporal modeling and incremental learning are important technical directions.

Combining lightweight models such as MobileNet with recurrent neural networks (RNNs) or temporal convolutional networks (TCNs) can capture sequences of food state changes and improve adaptability to dynamic scenarios. For example, in food and beverage monitoring, LSTM is used to temporal fusion of continuous frames to reduce false positives caused by occlusion or motion blur [13].

The incremental learning framework allows the model to continuously learn new categories or scenarios after deployment, avoiding full model retraining and adapting to dynamic changes in food types. For example, a lightweight incremental learning model based on EWC (Elastic Weight Consolidation) can achieve low-energy model updates on edge devices [19].

Multi-task learning architectures (such as simultaneous food classification and calorie estimation) reduce duplicate calculations and improve comprehensive processing efficiency in dynamic scenarios by sharing feature extraction layers [20].

2.5 Common problems of prior art

2.5.1 . The inherent contradiction between precision and efficiency is difficult to eradicate

Current lightweight models (such as MobileNet and ShuffleNet series) greatly reduce computational complexity and parameter volume through deep separable convolution, packet convolution and other technologies, laying the foundation for mobile deployment. However, this increase in efficiency often comes at the expense of some model characterization capabilities, resulting in significantly lower recognition accuracy than large models (such as ResNet-152) when dealing with extremely fine interclass differences (e.g., apple varieties of different varieties) or highly complex background interference. Although the Pareto optimal architecture can be automatically found through neural architecture search (NAS), the search process is computationally expensive, and the final model structure is complex, poorly interpretable, and difficult to popularize.

2.5.2 . The generalization ability and scenario adaptability of the model are insufficient

The generalization ability and scenario adaptability of the

ISSN 2959-6157

model are insufficient.

Most existing models are trained and optimized on specific, limited databases, with significant "data bias" issues. When the model is deployed to a new environment (e.g., different restaurant lighting, unknown cutlery styles) or faced with long-tail categories that don't appear in the training set, such as local snacks, its performance degrades significantly. Although data augmentation and transfer learning are widely used, they often fail to cover the infinite visual changes in the real world. The model lacks effective integration and utilization of human prior knowledge, making it difficult for it to respond to unknown scenarios through logical reasoning like humans.

2.5.3 . Robustness for dynamic and extreme scenarios is still lacking

Existing technologies have achieved certain results for a single type of interference (such as uniform lighting changes, slight occlusion), but it is still inadequate for extreme dynamic scenarios with multiple interferences and rapid changes. For example, the recognition rate of the model will drop dramatically when the light flickers violently, the food is completely obscured, or the liquid food (e.g., soup, sauce) is deformed. Current methods focus on the spatial feature extraction of static images, while the use of temporal information and contextual correlation in video streams is insufficient, and the modeling and prediction capabilities of scene dynamic evolution are lacking.

2.5.4 . There is a disconnect between the evaluation system and actual needs

The evaluation criteria of the current study generally rely on the top-1/top-5 classification accuracy on the closed test set, which does not fully reflect the performance of the model in real applications. A model that scores high on the test set may not be practical in real-world scenarios due to high response latency, high energy consumption, and unfriendly user interaction. The lack of comprehensive evaluation benchmarks covering multiple dimensions such as accuracy, speed, power consumption, and user experience hinders the transformation of research results into industrial applications.

2.5.5 . Strong hardware dependence and high deployment threshold

The optimal performance of many lightweight models is highly dependent on specific hardware and inference engines (e.g., specific models of GPUs, NPUs). A model that is efficient on one hardware may be inefficient on another. This strong hardware dependency raises the threshold and cost of technology implementation and large-scale deployment. Developers need to perform tedious adaptation and optimization for different terminal devices, which increases the complexity of technology applications

3. Conclusion

In summary, this review has examined the key challenges and recent advances in food image recognition under complex scenarios, including environmental disturbances, structural variations, and system constraints. While methods such as lightweight CNNs, attention mechanisms, and multimodal fusion have improved robustness and efficiency, several limitations remain—such as the trade-off between accuracy and speed, limited generalization, and inadequate performance in dynamic or extreme conditions. Future work should focus on enhancing model adaptability, integrating reasoning with perception, and developing comprehensive evaluation benchmarks. By addressing these issues, food recognition technology can evolve from laboratory settings to real-world applications, ultimately supporting smarter health, retail, and agricultural systems.

References

[1] Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans Neural Netw Learn Syst*. 2022;33(12):6999–7019. doi:10.1109/TNNLS.2021.3084827.

[2] Guo M-H, et al. Attention mechanisms in computer vision: a survey. *Comput Vis Media*. 2022;8(3):331–368. doi:10.1007/s41095-022-0271-y.

[3] Zhang Y, Yang Q. An overview of multi-task learning. *Natl Sci Rev*. 2018;5(1):30–43. doi:10.1093/nsr/nwx105.

[4] Zhang C, Yang Z, He X, Deng L. Multimodal intelligence: representation learning, information fusion, and applications. *IEEE J Sel Top Signal Process*. 2020;14(3):478–493. doi:10.1109/JSTSP.2020.2987728.

[5] Zhao D, Wu R, Liu X, Li Y. Localization of apple picking robot in complex background based on YOLO deep convolutional neural network. *Trans Chin Soc Agric Eng*. 2019;35(3):164–173. Available from: CNKI:SUN:NYGU.0.2019-03-021.

[6] Zhao X, Liu P, Tang X, Liu Y. A background modeling and object detection method adaptive to outdoor illumination changes. *Acta Autom Sin*. 2011;37(8):915–922. Available from: CNKI:SUN:MOTO.0.2011-08-004.

[7] Su H, Zhou J, Zhang Z. A survey on super-resolution image reconstruction. *Acta Autom Sin*. 2013;39(8):1202–1213. Available from: CNKI:SUN:MOTO.0.2013-08-005.

[8] Zeng Q, Chen Y, Wang Y, Liu J. A fast large-view image matching algorithm based on ORB. *Control Decis*. 2017;32(12):2233–2239. doi:10.13195/j.kzyjc.2016.1521.

[9] Li X, Liang R. A survey on occluded face recognition: from subspace regression to deep learning. *Chin J Comput*.

2018;41(1):177–207.
[10] Liang H, Wen X, Liang D, Zhang L. Fine-grained food image recognition using multi-level convolutional feature pyramid. *J Image Graph*. 2019;24(6):870–881.
[11] Liang S, Gu Y. A coarse-to-fine feature aggregation neural network with a boundary-aware module for accurate food recognition. *Foods*. 2025;14(3):383. doi:10.3390/foods14030383.
[12] Sarafis I, Papadopoulos A, Delopoulos A. Weakly supervised food image segmentation using vision transformers and segment anything model. *arXiv preprint* arXiv:2509.19028. 2025.
[13] Mijwil MM, Doshi R, Hiran KK, Al-Mistarehi AA. MobileNetV1-based deep learning model for accurate brain tumor classification. *Mesopotam J Comput Sci*. 2023;2023:29–38.
[14] Ma N, Zhang X, Zheng H-T, Sun J. ShuffleNet v2: practical guidelines for efficient CNN architecture design. In: *Proc Eur Conf Comput Vis (ECCV)*; 2018. p.116–131.
[15] Redmon J, Farhadi A. YOLOv3: an incremental improvement. *arXiv preprint* arXiv:1804.02767. 2018.
[16] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv preprint* arXiv:1602.07360. 2016.
[17] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: inverted residuals and linear bottlenecks. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*; 2018. p.4510–4520.
[18] Gao H, Tian Y, Xu F, Zhang L. A survey of model compression and acceleration for deep networks. *J Softw*. 2021;32(1):68–92. doi:10.13328/j.cnki.jos.006096.
[19] Wu Y, Chen Y, Wang L, Ye Y, Liu Z, Guo Y. Large scale incremental learning. In: *Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR)*; 2019. p.374–382.
[20] Zhang Y, Yang Q. A survey on multi-task learning. *IEEE Trans Knowl Data Eng*. 2021;34(12):5586–5609.