



Comparative Analysis of State-of-the-Art Deep Learning Models for Food Image Recognition

Prepared By

Name: Atheer Mohammed Al-Shahrani

ID No:252000400

Submitted in fulfillment of the requirements for the degree Master of Science
in Artificial Intelligence

At

College of Informatics
Midocean University

UAE

Supervisor Name: Dr. Yasser Abdel-Hamid

College of Informatics, Midocean University, Comoros

2026

Acknowledgements

At the beginning of this work, I thank God Almighty for the good ending and for enabling me to complete this research. Without His grace and guidance, this work would not have been possible.

I am also deeply thankful to Dr. Yasser Abdel-Hamid for his exceptional supervision, insightful guidance, and continuous support during the development of this study.

Additionally, I extend my heartfelt appreciation to my family and friends for their unwavering encouragement, patience, and moral support. Their presence and faith in me have been an invaluable source of strength throughout this academic journey.

Abstract

Food image recognition has become a major research topic in computer vision due to the increasing demand for applications such as automated meal recording, diet tracking, calorie estimation, and smart food systems. Despite significant advances in deep learning techniques, accurate food recognition remains a challenging task due to the considerable variations within a single category and the significant visual similarities between different food categories.

This work presents a systematic comparison of the latest deep learning models for food image recognition using three publicly available datasets of increasing complexity: Food-11, Food-41, and Food-101. The selected models were evaluated within a single experimental framework based on their classification accuracy, their ability to adapt to visual variation, their generalizability to datasets of varying difficulty, and their performance in accurate food classification tasks. The results highlight the advantages and disadvantages of each model.

Table of Contents

Chapter 1	1
1.1 Problem Definition.....	1
1.2 Thesis Objectives	2
1.3 Thesis Contribution	3
1.4 Thesis Organization.....	4
Chapter 2	5
2.1 Introduction	5
2.2 Models Used in Food Image Recognition	5
2.2.1 Convolutional Neural Networks (CNNs).....	5
2.2.2 Deep CNN Structures	6
2.2.3 CNN Models with Attention	6
2.2.4 Models of State Space.....	7
2.3 Food Image Datasets.....	7
Chapter 3	9
3.1 Introduction	9
3.2 Traditional Approaches to Food Image Recognition	9
3.3 Transition to Artificial Intelligence–Based Methods	10
3.4 Contributions of Artificial Intelligence to Food Image Recognition.....	10
3.5 Comparative Summary of Existing Studies	11
Table 3.1: Summary of Key Deep Learning Models for Food Recognition	11
3.6 CNN-Based Models in Previous Studies.....	12
3.7 Models that Improve Attention	12
3.8 Object Detection–Based Approaches	13
3.9 State-Space Models	13
References.....	14

List of Abbreviations

- **Acc:** Accuracy
- **Adam:** Adaptive Moment Estimation
- **AI:** Artificial Intelligence
- **ANN:** Artificial Neural Network
- **AUC:** Area Under the Curve
- **CNN:** Convolutional Neural Network
- **DL:** Deep Learning
- **F1-score:** F1 Measure
- **Food-11:** Food-11 Image Dataset
- **Food-41:** Food-41 Image Dataset
- **Food-101:** Food-101 Image Dataset
- **GPU:** Graphics Processing Unit
- **LR:** Learning Rate
- **ML:** Machine Learning
- **mAP:** Mean Average Precision
- **Prec:** Precision
- **Rec:** Recall
- **Res-VMamba:** Residual Vision Mamba
- **SGD:** Stochastic Gradient Descent
- **SHAP:** SHapley Additive exPlanations
- **SSM:** State-Space Model
- **TL:** Transfer Learning
- **ViT:** Vision Transformer
- **VMamba:** Vision Mamba
- **XAI:** Explainable Artificial Intelligence

Figure 1. Overview of the thesis structure developed in the current semester. 4

Chapter 1

Introduction

1.1 Problem Definition

Food image recognition is a crucial field, supporting nutrition and health applications in various ways, including calorie counting, nutritional tracking, diet evaluation, and the development of smart eating plans. Despite the success of deep learning models, the complexity and diversity of food data make it challenging to identify food types from images. Even within the same category, food images vary considerably. For example, preparation and presentation, lighting, and camera angle all influence their appearance.

Furthermore, different food types often share similar appearances due to common characteristics such as color, texture, and overall shape. Because these characteristics are sensitive, classifying food images using traditional algorithms is difficult.

While numerous deep learning models have been proposed for food image recognition, most previous studies have focused on the architecture of these models using small datasets. This approach fails to provide a comprehensive picture of the models' reliability, scalability, and efficiency across a wide range of scenarios and datasets. Consequently, a systematic comparative study using a consistent evaluation methodology for various deep learning models is needed.

1.2 Thesis Objectives

This research primarily aims to conduct a comprehensive evaluation of the latest deep learning models for food image identification using increasingly complex datasets.

The objectives of this research are as follows:

1. To evaluate the effectiveness of a representational convolutional neural network (CNN) in food image classification as a basic methodology.
2. To investigate the effectiveness of an attention-enhanced convolutional neural network.
3. To improve the quality and accuracy of food image classification and determine the ability of state-space models to find general and local visual associations in tasks such as food image identification.
4. To measure the model's efficiency on the following three databases: Food-11, Food-41, and Food-101.
5. Measurement tests will be used on a dataset categorized by difficulty level to determine the quality of work performed by different groups and how performance varies each time.

1.3 Thesis Contribution

This thesis makes important progress in the field of food image recognition:

- **A unified benchmarking framework:** A unified experimental benchmark framework is defined for evaluating different deep learning architectures under several similar training and evaluation conditions. This provides a fair and logical comparison between them.
- **Comparing the analysis of different structures:** To better understand all the advantages and disadvantages of each structure, this method compares convolutional neural network models, attention-enhanced models, and state-space models by testing them on datasets of varying levels of complexity.
- **Accurate classification analysis:** Particularly for visually similar food categories, the study provides a quantitative and qualitative assessment of the impact of architectural design choices on food classification accuracy.
- **The impact of dataset complexity:** This illustrates how dataset complexity affects the success of scaling and applying a model to new cases. This provides us with more realistic data that can help us choose datasets and build models in the future.
- **In practical use:** The results allow us to choose the deep learning models that are best used, such as how to assess nutrition, estimate calories, and design smart diets.

1.4 Thesis Organization

Overall Thesis Organization

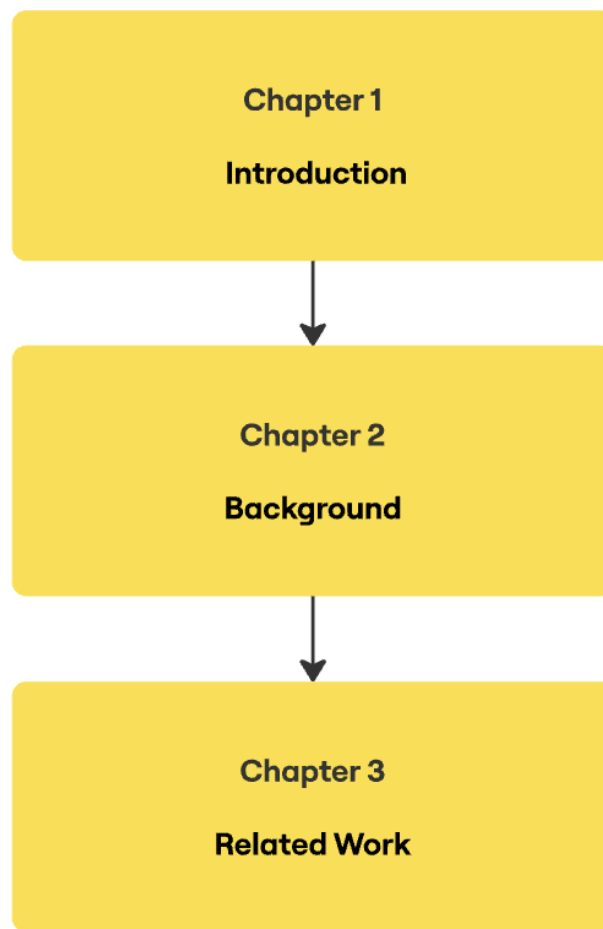


Figure 1. Overview of the thesis structure developed in the current semester.

Chapter 2

Background

2.1 Introduction

This is why food imagery is a crucial area within computer vision, where it competes for the automatic classification of food items from images. This functionality has broad applications in complex food systems, food monitoring, diet planning, and calorie estimation.

The increased availability of visual data and advancements in field technology have accelerated research in this area. Deep learning architectures have replaced older algorithms, relying on features from renowned food image experts.

This guide provides essential information and discusses common fundamental analyses used in food image analysis. Each new model is designed to offer a comprehensive understanding of its workings before being explored in new applications

2.2 Models Used in Food Image Recognition

This section discusses the main categories of deep learning models used to identify food images. Each category of models focuses on a specific type of structure.

2.2.1 Convolutional Neural Networks (CNNs)

are the basic deep learning models utilized for tasks like food photo identification. Convolutional layers, activation functions, and pooling operations enable CNNs to directly learn visual features from an image's raw pixels.

CNNs learn hierarchical representations to identify food images. Basic visual patterns like edges and textures are learned by the earliest layers. More intricate semantic elements pertaining to food presentation and structure are learned at deeper levels.

Because of this hierarchical feature learning, CNNs can adapt to changes in lighting, background noise, and food presentation. CNNs are the main component of most contemporary food photo identification systems and do not require human feature engineering.

2.2.2 Deep CNN Structures

Deeper CNN architectures were created to improve recognition. Models that enhance simple CNNs by incorporating more architectural components and deepening the network include VGG, ResNet, DenseNet, and Inception. This allows features to be used once more and strengthens learning.

Finding more intricate and detailed visual patterns is made easier by these designs, which is essential for distinguishing between seemingly identical food categories. Dense connectivity and residual connections are two strategies to facilitate deep network training and accelerate convergence.

In food image recognition, deep CNN architectures are commonly used as foundational networks and frequently paired with transfer learning methods.

2.2.3 CNN Models with Attention

To better focus on the most valuable parts of an image, attention-enhanced CNN models apply attention techniques to standard CNN designs. Different parts of space are handled differently by attention processes. Rather, they direct the network to important locations such as components, textures, and structural elements.

When it comes to identifying food images, attention processes help eliminate background information that is not required. They also make it easier to distinguish between food groups that appear to be similar. To make significant visual cues stand out, these models typically work by modifying the weights of feature maps or merging features from several levels.

CNN models with attention work well for applications that require very precise food identification, where even slight alterations in the image are very important.

2.2.4 Models of State Space

State-space models provide a fresh perspective on the operation of computer vision. While CNNs primarily focus on local spatial features, state-space models look for long-range connections and global context in images.

State-space models for food picture identification are made to efficiently capture both more general contextual interactions and localized visual features. These models make it simpler to describe complex food images while maintaining computational efficiency by combining state-space modeling with deep learning frameworks.

Although state-space models are still relatively new in this field, they appear to be a good starting point for future research on food image identification.

2.3 Food Image Datasets

Libraries of food photos are crucial for developing and testing recognition algorithms. These datasets differ in terms of the quantity of food categories, the variety of images, the quality of the images, and the precision of the annotations.

The way that models learn and generalize is directly impacted by the characteristics of a dataset. Food databases can be small, with only a few categories, or large, with a wide variety of foods and difficult-to-read labels. Knowing how difficult a dataset is is essential to selecting the best model architectures and comprehending trial results.

Chapter 3

Related Work

3.1 Introduction

Due to the rapid development of visual data and the proliferation of mobile devices, automatic food detection has become a major research topic in computer vision and artificial intelligence. Food image recognition has attracted considerable attention due to its importance in applications such as nutritional assessment, health monitoring, calorie estimation, and advanced food-related systems.

This chapter reviews previous studies on food image identification, focusing on the evolution of food recognition methods over time. It begins with traditional computer vision techniques, then moves on to artificial intelligence-based techniques and their capabilities. A table comparing the different models used in previous research is then presented, followed by a detailed explanation of each model.

3.2 Traditional Approaches to Food Image Recognition

Before the advent and development of artificial intelligence, most food image recognition systems relied on traditional computer vision techniques. They used manually designed visual methods, such as color mapping, texture description, and shape information. Later, they employed machine learning techniques, such as support vector machines (SVMs) and nearest-neighbor algorithms (k-NNs), to rank the features they were able to find.

These methods proved effective in controlled environments, but they failed to prove effective on real food images. Changes in lighting, background blurring, obscuring, and significant variations within the

same category all make it difficult to distinguish between different foods. Furthermore, using specifically designed features requires a deep understanding of the subject matter, which makes development challenging as the number of food categories increases.

3.3 Transition to Artificial Intelligence–Based Methods

The limitations of traditional feature-based methods led to a shift towards AI-powered solutions. Coinciding with the availability of massive datasets of food images and rapid advancements in computing, researchers were encouraged to apply deep learning techniques to overcome the challenges of traditional models.

AI techniques enable models to learn directly from raw visual data, unlike older methods. This helps them learn more effectively and distinguish between different visual representations. This upgrade represented a significant leap forward in food image recognition, reducing the likelihood of system crashes when conditions change and enabling them to work with additional types of datasets.

3.4 Contributions of Artificial Intelligence to Food Image Recognition

Artificial intelligence (AI) has revolutionized food photo identification by integrating machine learning with advanced object representation capabilities. Early AI models, such as convolutional neural networks (CNNs), were able to learn hierarchical visual information, resulting in significant improvements in object recognition accuracy.

Later research improved CNN-based models by adding deeper topologies, attention processes, and other ways of modeling. These changes made identification algorithms stronger at sorting food into

small groups, working with difficult backgrounds, and handling huge datasets. In summary, AI turned food photo recognition from a hard task with hand-made characteristics into a powerful, data-driven task.

3.5 Comparative Summary of Existing Studies

Table 3.1 lists some major deep learning models that may be used to compare studies on identifying food pictures. The table displays the model architectures used in previous research, the datasets used for testing, the performance metrics reported, and the main benefits and downsides of each strategy.

Table 3.1: Summary of Key Deep Learning Models for Food Recognition

Study	Model	Architecture Type	Dataset	Performance	Key Strengths	Key Limitations
Liu et al. (2016)	GoogLeNet (Inception)	CNN	Food-101, UEC-256	Food-101 Top-1 $\approx 77\%$	Established a strong deep-learning baseline; significant improvement over handcrafted features	Limited ability in fine-grained recognition
Boyd et al. (2024)	GoogLeNet (Inception)	CNN	Custom 20-class dataset	Val Acc: 79% , Train Acc:	Excellent feature reuse; stable convergence	Performance drops on large-scale complex datasets
Chen et al. (2025)	MAF-Net	Attention-Enhanced CNN	Food-101, Vireo-172, UEC-100	Food-101: 90.22%	Strong fine-grained recognition; discriminative attention mechanisms	Computationally heavy due to multi-level fusion
Karabay et al. (2024)	YOLOv8-XL	Object Detection	CAFSD	mAP50 = 0.677	Strong fine-grained recognition; discriminative attention mechanisms	Not designed for classification; detection-only
Chen et al. (2024)	Res-VMamba	State-Space Model	CNFOOD-241	Top1: 79.54%	Efficient global–local modeling; competitive with ViT	Limited pretrained weights; new architecture
Bossard et al. (2014)	Standard CNN	CNN	Food-101	Baseline $\approx 56\%$	Introduced a foundational dataset for food recognition	High noise; strong intra-class variation

Islam et al. (2020)	AlexNet, VGG, ResNet	CNN	Food-5K, Food-11	Food-11 $\approx$ 83%	Fast training; good for low-variation datasets	Accuracy decreases as class count increases
---------------------	----------------------	-----	------------------	------------------------------	--	---

3.6 CNN-Based Models in Previous Studies

Most AI-powered systems that identify food pictures use CNN-based models. Bossard et al. (2014) published the Food-101 dataset with standard CNN baselines, making it one of the first large-scale tests for identifying food and revealing how hard it is to work with noisy labels and a lot of variety between classes.

Liu et al. (2016) shown that deeper CNN architectures, such as GoogLeNet (Inception), significantly outperform conventional methods that rely on handcrafted features. This makes CNNs strong base models. Islam et al. (2020) evaluated CNN architectures, such as AlexNet, VGG, and ResNet, using datasets like Food-11, indicating high accuracy for datasets of modest complexity. But their data showed that performance is worse as the number of food categories increases.

Boyd et al. (2024) did a more in-depth investigation of CNN-based models and showed that deeper architectures had steady convergence and could reuse features well. Even with these improvements, CNN-based approaches still have problems distinguishing food items in detail when the differences between categories are minimal.

3.7 Models that Improve Attention

To help CNNs detect the difference between pictures better, attention-enhanced models were made that focus on the areas of the picture that are most beneficial. Attention processes assist models in paying more attention to important visual cues, such as ingredients and texture patterns, and less attention to background information that isn't important.

Chen et al. (2025) proposed MAF-Net, a CNN architecture that enhances attention through the integration of multi-level feature fusion and attention refinement. Their testing indicated that their way of noticing little features in food performed better. Still, attention-based models are sometimes harder to calculate, which might make them less helpful in big scenarios.

3.8 Object Detection–Based Approaches

Some studies examined frameworks for identifying items in food-related activities. Karabay et al. (2024) employed YOLOv8-XL to recognize food items in complex settings, achieving notable detection efficacy.

Detection-based approaches are good for discovering food, but they shouldn't be used to sort things. This means that they often take additional steps to process, which makes them less effective for immediately sorting food pictures.

3.9 State-Space Models

State-space models are a new field of research that tries to describe long-range interdependence and the global context in a useful way. Chen et al. (2024) introduced Res-VMamba, which combines residual learning with chosen state-space modeling.

Tests of CNFOOD-241 demonstrated that it worked just as well as CNN-based approaches. State-space models are still not very well recognized for identifying food pictures because they are new in terms of architecture, and there aren't many pretrained models accessible.

References

Becker, D. (n.d.). Food-101 [Data set].

Kaggle. <https://www.kaggle.com/datasets/dansbecker/food-101>

Chen, C.-S., Chen, G.-Y., Yang, P.-Y., & Chen, D. (2024). Res-VMamba: Fine-grained food category visual classification using selective state space models with deep residual learning. arXiv. <https://arxiv.org/abs/2402.15761>

Chen, Z., Wang, J., & Wang, Y. (2025). Enhancing food image recognition by multi-level fusion and the attention mechanism. *Foods*, 14(3), 461. <https://www.mdpi.com/2304-8158/14/3/461>

K-mader. (n.d.). Food41 [Data set].

Kaggle. <https://www.kaggle.com/datasets/kmader/food41>

Karabay, A., Varol, H. A., & Chan, M. Y. (2025). Improved food image recognition by leveraging deep learning and data-driven methods with an application to Central Asian Food Scene. *Scientific Reports*, 15(1), 14043. <https://www.nature.com/articles/s41598-025-95770-9>

Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., & Ma, Y. (2016). DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment. arXiv. <https://arxiv.org/abs/1606.05675>

Liu, D., Zuo, E., Wang, D., He, L., Dong, L., & Lu, X. (2025). Deep learning in food image recognition: A comprehensive review. *Applied Sciences*, 15(14), 7626. <https://www.mdpi.com/2076-3417/15/14/7626>

Liu, G., Jiao, Y., Chen, J., Zhu, B., & Jiang, Y.-G. (2024). From canteen food to daily meals: Generalizing food recognition to more practical scenarios. arXiv. <https://arxiv.org/abs/2405.16478>

Min, W., Jiang, S., Liu, L., & Rui, Y. (2019). A survey on food computing: An overview of recent advances and challenges. *ACM Computing Surveys (CSUR)*, 52(3), 1–36. <https://doi.org/10.1145/3329168>

Min, W., Liu, L., Luo, Z., & Jiang, S. (2019). Ingredient-guided cascaded multi-attention network for food recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3), 1–22. <https://dl.acm.org/doi/epdf/10.1145/3329168>

Min, W., Wang, Z., Liu, Y., Luo, Z., Kang, L., Wei, X., ... & Jiang, S. (2023). Large-scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8), 9939-9955. <https://ieeexplore.ieee.org/document/9793740>

Moumane, K., & El Biki, S. (2023). Food recognition and nutrition estimation using MobileNetV2 CNN architecture and transfer learning. *Procedia Computer Science*, 230, 1785–1794. <https://www.sciencedirect.com/science/article/pii/S1566253523001756>

Trolukovich, P. (n.d.). Food-11 image dataset [Data set]. Kaggle. <https://www.kaggle.com/datasets/trolukovich/food11-image-dataset>

Xiao, J., et al. (2023). ISIA Food-500: A huge-scale benchmark for food recognition. OpenReview. <https://openreview.net/forum?id=YicbFdNTTy>

Zhao, Y., Feng, S., Zhang, X., & Hou, Y. (2024). A novel multi-level feature fusion network for fine-grained food image recognition. *Informatics*, 10(6), 126. <https://www.mdpi.com/2313-433X/10/6/126>