



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

First and foremost, I would like to express my sincere gratitude to Allah Almighty for granting me the strength, clarity, and perseverance necessary to complete this research. Without His guidance and blessings, 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. His expertise, constructive feedback, and academic dedication have greatly influenced both the quality and direction of this research.

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 emerged as a central task in food computing, supporting applications such as automated meal logging, dietary monitoring, restaurant menu indexing, and intelligent food systems. Recent advancements in deep learning have introduced diverse architectures—ranging from classical Convolutional Neural Networks (CNNs) to attention-based models, Vision Transformers (ViTs), and modern state-space architectures such as Mamba—each offering unique capabilities for fine-grained visual classification.

This research presents a systematic comparative analysis of state-of-the-art deep learning models for food image recognition using three public datasets of varying complexity: Food-11, Food-41, and Food-101. The selected models are evaluated based on classification accuracy, robustness to intra-class variation, generalization ability across difficulty levels, and performance on fine-grained recognition tasks.

The findings provide a unified benchmark that highlights architectural strengths and limitations across different model families. This study offers practical guidance for selecting effective models in real-world food recognition applications and identifies future research directions for improving fine-grained visual understanding within food computing.

Keywords:

Deep Learning; Food Image Recognition; Convolutional Neural Networks (CNNs); Attention Mechanisms; Vision Transformers (ViT); State-Space Models (Mamba); Fine-Grained Classification; Comparative Study; Food-11; Food-41; Food-101.

Table of Contents

Chapter 1.....	1
Introduction	1
1.1 Problem Definition.....	1
1.2 Thesis Objectives	1
1.3 Thesis Contribution	2
1.4 Thesis Organization.....	3
Chapter 2.....	4
2.1 Introduction to Food Image Recognition	4
2.2 CNN-Based Food Recognition Methods.....	4
2.3 Attention-Enhanced CNN Approaches	5
2.4 State-Space Models for Food Image Recognition.....	5
2.5 Food Image Datasets.....	5
2.6 Quantitative Results Summary	6
2.7 Research Gaps and Opportunities.....	7
2.8 Conclusion	7
Chapter 3.....	8
3.1 Introduction	8
3.2 History.....	8
3.3 Summary	9
References.....	10

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. 3

Chapter 1

Introduction

1.1 Problem Definition

Food image recognition is very important for current uses including calorie counting, nutrition monitoring, dietary evaluation, and smart food-related systems. Even with recent breakthroughs in deep learning, it is still hard to reliably identify food items from pictures since food visual data is so complex.

Food photographs frequently demonstrate substantial intra-class diversity, since identical food items can seem markedly different based on cooking methods, presentation, lighting circumstances, and perspective. Inter-class resemblance is also widespread, as food groups that look different may have similar textures, colors, or patterns in their structures. These traits make it hard to handle food picture identification using standard methods since it is a fine-grained classification challenge.

Many deep learning models have been suggested for recognizing food images, however much of the research so far has only looked at one type of architecture and one dataset. This fragmented assessment technique makes it hard to make solid judgments about how well the model works, how well it scales, and how well it generalizes across datasets with different levels of complexity. So, there is an obvious need for a single, organized comparison study that tests multiple deep learning architectures in the same way.

1.2 Thesis Objectives

The primary aim of this thesis is to do an exhaustive comparative examination of cutting-edge deep learning models for food picture identification across datasets of escalating complexity.

The specific objectives of this study are:

1. To assess the efficacy of representative convolutional neural network (CNN) models as foundational methodologies for food picture categorization.
2. To examine the efficacy of attention-enhanced CNN architectures in enhancing fine-grained food recognition.
3. To look at how well state-space models can find global and local visual dependencies for tasks that include recognizing food images.
4. To evaluate model performance and generalization across the Food-11, Food-41, and Food-101 datasets.

5. To use standard assessment measures to examine how well different classifications work and to look at how performance changes as the dataset gets more complicated.

1.3 Thesis Contribution

This thesis makes several important contributions to the field of food image recognition:

- **Unified Benchmarking Framework:**
It sets up a single experimental framework for testing different deep learning architectures under the same training and assessment settings, which makes it possible to compare them fairly and systematically.
- **Cross-Architecture Comparative Analysis:**
It offers a systematic comparison of CNN-based, attention-enhanced, and state-space models across datasets with different levels of complexity, giving us fresh information about their strengths and weaknesses.
- **Fine-Grained Recognition Analysis:**
The study provides both quantitative and qualitative examination of the impact of architectural design decisions on fine-grained food categorization performance, especially for visually comparable food categories.
- **Dataset Complexity Impact:**
It shows how the complexity of a dataset affects the scalability and generalization of a model, providing empirical information to help with future dataset selection and model design.
- **Practical Applicability:**
The results support making smart choices about which deep learning models to use in real-world situations like analyzing nutrition, estimating calories, and developing smart food systems.

1.4 Thesis Organization

Overall Thesis Organization

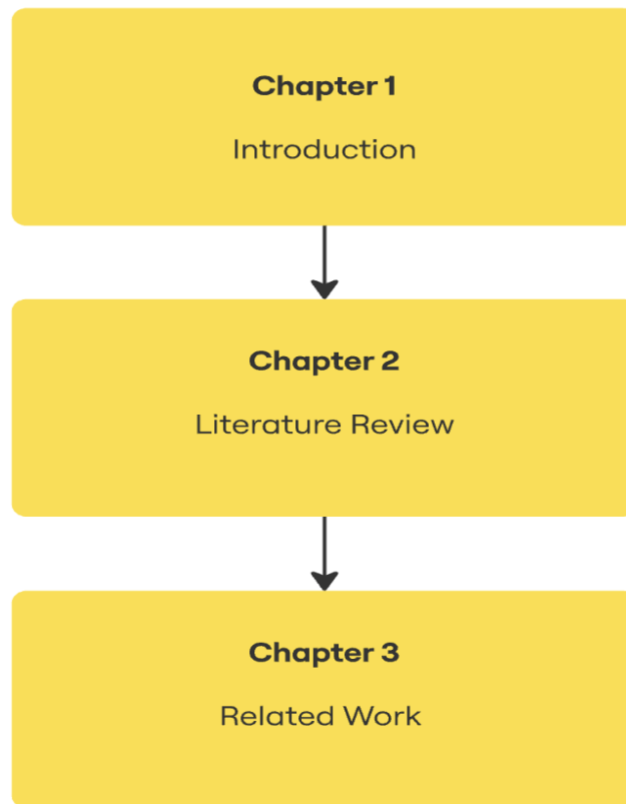


Figure 1. Overview of the thesis structure developed in the current semester.

Chapter 2

2.1 Introduction to Food Image Recognition

Food image recognition has become a significant field of study since it is becoming more essential in areas like dietary evaluation, calorie calculation, nutrition monitoring, and smart food-related systems. In recent years, automatic detection of food products from photographs has become quite popular. This is because more and more visual food data is becoming available and computing resources are getting better.

Early research included conventional machine learning methods integrated with manually produced visual attributes. But these methods had a hard time dealing with the great unpredictability and complexity that come with food imagery. Because of this, deep learning methods, especially convolutional neural networks, have become the most popular way to recognize food images. This chapter discusses important changes in the literature, such as CNN-based models, attention-enhanced architectures, state-space models, widely used datasets, and reported quantitative results. It gives the work in this thesis a background.

2.2 CNN-Based Food Recognition Methods

Convolutional Neural Networks (CNNs) were among of the first deep learning models that operate well for recognizing food images. CNNs learned hierarchical feature representations directly from raw photos, which made them far better than previous hand-crafted feature methods, especially when there was background clutter, lighting changes, or different ways of presenting food.

Bossard et al. (2014) introduced the Food-101 dataset, which was a major step forward in this field. It established a difficult standard with a lot of diversity within classes and labels that weren't always clear. Liu et al. (2016) built on this dataset to create the DeepFood framework utilizing GoogLeNet (Inception). This showed that CNNs are powerful baseline models for food categorization tasks and that they function far better than traditional approaches.

Later studies looked at more complex and deeper CNN designs, such as VGG, ResNet, and DenseNet, with the goal of boosting recognition accuracy by adding more depth and reusing features. Boyd et al. (2024) performed a comparative analysis of several CNN models, indicating enhanced performance and consistent convergence for DenseNet-based designs. Even with these improvements, CNN-based models still have trouble with fine-grained categorization, especially when food categories that look identical only differ in very small ways.

2.3 Attention-Enhanced CNN Approaches

To get around the problems with regular CNNs, food recognition models now include attention processes. Attention-enhanced CNN architectures enable models to concentrate on salient regions within a picture, including components, textures, and structural patterns, while diminishing the impact of extraneous background information.

Chen et al. (2025) recently developed MAF-Net, an attention-enhanced CNN architecture that combines multi-level feature fusion with attention refinement techniques. Experimental results showed that the classification performance was better on benchmark datasets like Food-101, especially for jobs that required fine-grained recognition. These results show that attention processes can improve the capacity of CNN-based food recognition models to tell the difference between different foods, but this typically comes at the cost of making the models more complicated to use.

2.4 State-Space Models for Food Image Recognition

State-space models are a new way of designing computer vision systems that is different from the usual CNN and transformer-based methods. These models are intended to efficiently capture long-range relationships while yet being good for computing.

Chen et al. (2024) put up Res-VMamba, which combines deep residual learning with selective state-space modeling. Their findings on detailed food datasets like CNFOOD-241 showed that state-space models can work just as well as CNN-based models. Even though they show potential, state-space models are still not very well known in the field of food picture identification. This shows that further testing is needed on datasets of different levels of difficulty.

2.5 Food Image Datasets

Dataset features are very important for judging food image recognition algorithms. Food picture databases varies a lot in how many classes they have, how many different kinds of images they have, and how good their labels are. This has a direct effect on how well models work and how well they generalize.

This thesis centers on three extensively utilized datasets that exhibit escalating levels of complexity. Food-11 is a simple setting that may be used for baseline testing. Food-41 adds some moderate fine-grained variation, while Food-101 is a high-complexity benchmark with a lot of diversity between classes and noisy labeling. Previous research shows that model performance frequently becomes worse when the dataset gets more complicated. This shows how important it is to test models on more than one dataset.

2.6 Quantitative Results Summary

This part gives a short quantitative summary of several exemplary research on recognizing food images. Table 2.1 lists the most important deep learning models that have been written about in the literature, together with the datasets and performance measures that were used to test them. The table shows the numerous architectural techniques that have been used in the past and gives a quick look at the results that have been reported from different experimental situations.

Table 2.1: 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

2.7 Research Gaps and Opportunities

There are still several holes in the current research on food image identification, even though it has come a long way. The majority of research concentrate on a singular architectural family, hence constraining understanding of the comparative performance of many model types under uniform settings. Cross-dataset assessments are still not very common, even if there are big variances in how complicated and visually varied the datasets are.

Additionally, current research frequently presents findings based on isolated benchmarks, complicating the analysis of performance scalability and generalization trends. These constraints underscore the necessity for a cohesive comparison methodology that methodically assesses various deep learning architectures across datasets exhibiting differing degrees of complexity.

2.8 Conclusion

The literature examined in this chapter illustrates consistent progress in food image recognition, with CNN-based models establishing robust baselines, attention-enhanced architectures facilitating fine-grained classification, and novel state-space models presenting advantageous trade-offs between performance and efficiency. The absence of standardized assessments across architectures and datasets necessitates the comparison study presented in this thesis, which seeks to elucidate model behavior, scalability, and generalization performance.

Chapter 3

3.1 Introduction

Food image recognition has garnered heightened study interest owing to its significance in applications including dietary evaluation, health surveillance, calorie estimate, and intelligent food systems. As visual data grows quickly and mobile devices become more common, automatic food recognition has become a major research problem in computer vision and artificial intelligence.

Researchers have suggested a lot of different ways to do things throughout the years. These have changed from simple image processing methods to more complicated deep learning systems. This chapter looks at the most important events and technical advances in food picture identification. It also puts the current thesis in the context of other studies in the field. It shows how data-driven learning algorithms have slowly solved problems with visual variability, fine-grained classification, and dataset complexity.

3.2 History

Initial research in food picture identification predominantly utilized conventional computer vision techniques and manually engineered characteristics. These methods used color, texture, and form descriptors together with traditional classifiers like support vector machines and k-nearest neighbors. While these algorithms worked well in limited situations, they had trouble applying to real-world food photographs because of background noise, variations in lighting, and a lot of variety within each class.

As machine learning techniques got better, stronger classifiers were added to make recognition work better. But these models still relied significantly on characteristics that were developed by hand, which made it hard for them to work with a wide range of food types and datasets.

In the last ten years, there has been a big movement toward deep learning-based methods, especially convolutional neural networks (CNNs). CNNs made it possible to train hierarchical visual representations from raw photos all the way through, which greatly improved food identification performance and made deep learning the most popular way to achieve this.

More recently, research has concentrated on improving CNN architectures to more effectively tackle fine-grained food categorization difficulties. Attention methods were added to help models focus on important parts of images, including ingredients and texture patterns, while ignoring background information that isn't important.

Simultaneously, new architectural paradigms, such state-space models, have been looked into as possible replacements for traditional CNN and transformer-based methods. These models try to better capture long-range relationships and global context, which shows how deep learning approaches are always becoming better at finding more scalable and expressive ways to solve hard food recognition problems.

3.3 Summary

The development of food picture identification shows a clear shift from hand-made feature approaches to deep learning-based architectures that can learn rich and distinct visual representations. Improvements in CNNs, models that boost attention, and new state-space architectures have made recognition much better, especially for very specific food categories.

Nonetheless, current research frequently assesses these methodologies in isolation, without a comprehensive examination of their performance across datasets of differing complexities. This constraint necessitates the integrated comparison study delineated in the ensuing parts of this thesis.

References

- Becker, D. (n.d.). *Food-101* [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/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. Retrieved from <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>
- 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. Retrieved from <https://www.kaggle.com/datasets/trolukovich/food11-image-dataset>
- 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>