



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.

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

i

Chapter 1	1
1.1 Problem Definition	1
1.2 Thesis Objectives	2
1.3 Thesis Contribution	2
Chapter 2	3
2.1 Deep Learning for Food Image Recognition	3
2.2 Convolutional Neural Networks (CNNs)	4
2.3 Attention-Based CNN Models	4
2.4 Vision Transformers (ViTs)	5
2.5 State-Space Models (e.g., VMamba)	5
2.6 Food Image Datasets	6
2.7 Quantitative Results Summary	7
2.8 Summary of Literature Review	7
Chapter 3	8
3.1.1 CNN-Based Food Recognition Methods	9
3.1.2 Attention-Enhanced CNN Approaches	9
3.1.3 Transformer-Based Food Recognition	10
3.1.4 State-Space Models for Food Recognition	10
3.2 Multi-Food Detection and Dataset Challenges	10
3.3 Limitations in Existing Research	10
3.4 Summary	11
References	11

Chapter 1

Introduction

Deep learning has made a big difference in the field of food picture identification. It has made it possible to automatically analyze eating patterns, estimate calories, assess nutrition, and create smart food systems. Food photos frequently display significant intra-class diversity and nuanced inter-class similarity, necessitating the development of model's adept at discerning subtle visual cues.

Modern designs, such as CNNs, attention-based networks, Vision Transformers, and state-space models, all have their own ways of extracting features, and each has its own benefits. Nevertheless, despite swift advancements, scant research has offered a cohesive comparison of these model families utilizing datasets of diverse complexity.

This chapter explains why the study was done, what the problems were, what the thesis goals were, what the contributions were, and how the study was organized.

1.1 Problem Definition

There are still certain problems with current research:

- Most previous studies only look at one design on one dataset, which makes it hard to apply their findings to other situations.
- There are still not many comparisons between CNNs, attention-based architectures, Transformers, and state-space models, even though they all use different ways to encode features.
- The datasets from earlier research vary greatly in how hard they are to work with, which makes it hard to find consistent performance or strengths that are particular to certain architectures.
- Many works do not do a good job of dealing with fine-grained difficulties, such as food products that seem similar or different ways of cooking.

This results in a significant deficiency:

There exists no exhaustive comparison research that systematically assesses contemporary deep learning architectures for food picture identification across datasets varying in complexity and granularity.

1.2 Thesis Objectives

The primary aim of this thesis is to execute an exhaustive comparative examination of cutting-edge deep learning models for food picture recognition using the **Food-11**, Food-41, and Food-101 datasets.

To do this, the project seeks to:

- Assess the efficacy of four principal architectural families: CNNs, attention-enhanced CNNs, Vision Transformers, and state-space (Mamba) models.
- Evaluate how the model behaves at different levels of difficulty.
- Look at how strong, general, and detailed the categorization abilities are.
- Find architectural characteristics that have a big effect on how well recognition works.

1.3 Thesis Contribution

This research offers the subsequent contributions:

1. A single benchmarking framework that looks at how well different deep learning architectures operate on several food picture datasets.
2. A thorough assessment of accuracy, robustness, fine-grained categorization, and generalization.
3. An architectural analysis that shows how convolution, attention, transformers, and state-space techniques affect performance.
4. Useful tips for choosing the right models for nutrition analysis, meal tracking apps, and smart food systems.
5. A better knowledge of how the complexity of a dataset affects the performance of a model, which will help in the creation of new datasets and models in the future.

Chapter 2

Overview

This chapter provides a comprehensive analysis of current research on food image recognition employing advanced deep learning architectures. It methodically explores the progression of models, ranging from initial CNN-based frameworks to attention-enhanced networks, Vision Transformers (ViTs), and state-space architectures. The review also emphasizes significant datasets frequently utilized in food computing and concludes with a quantitative comparison that encapsulates findings from representative studies.

2.1 Deep Learning for Food Image Recognition

Food picture recognition is a key part of food computing. It helps with things like figuring out how many calories are in a meal, checking your diet, smart kitchen systems, and keeping track of your nutrition. The work is still hard because of:

- High variety within classes (for example, the same food made or served in different ways).
- A lot of resemblance between classes (for example, meals that look same).
- Complicated backdrops, changing lighting, and anything that get in the way.
- Fine-grained differences that are caused by the ingredients or the way the food is cooked.

Traditional machine-learning approaches had a hard time with these complications since they relied on hand-made characteristics.

Deep learning, especially Convolutional Neural Networks (CNNs), changed the discipline by letting models learn hierarchical, discriminative features directly from raw pictures.

As time has gone on, more complex architectures have come forth, such attention-enhanced CNNs, Vision Transformers (ViTs), and state-space models like VMamba. Every family adds new ways to fix problems with older designs and make fine-grained recognition work better.

2.2 Convolutional Neural Networks (CNNs)

CNNs were the first step toward contemporary food identification. Early designs like GoogLeNet/Inception, VGG, and ResNet always did well on benchmark datasets.

Deep Food (Liu et al., 2016) was one of the first important efforts to use GoogLeNet to sort food items from Food-101 and UEC256. It did better than previous hand-crafted methods.

More research kept improving CNN models. **Boyd et al. (2024)** assessed several CNN variations and determined Dense Net to be the superior design following comprehensive parameter tuning. Their improved Dense Net got a high validation accuracy on a 20-class dataset with fine-grained details.

Strengths of CNNs:

- Good at finding hierarchical visual patterns
- Works well for classifying big amounts of data
- Stable and efficient in terms of computation

Limitations:

- Limited ability to simulate long-range dependencies
- Hard to tell the difference between very identical meals
- Dependence on local receptive fields

These constraints prompted the incorporation of attention techniques to improve feature discrimination.

2.3 Attention-Based CNN Models

Attention mechanisms boost CNNs by enabling models to highlight the most informative picture areas. This is particularly important in food identification, where minor visual cues—such as texture, garnish, or ingredients—are critical for fine-grained categorization.

A notable instance is MAF-Net (**Chen et al., 2025**), which integrates:

- Combining features at different levels
- Self-attention refinement modules
- Maximizing KL-divergence
- Subclass Center Loss to make fine-grained separability better

MAF-Net got more over 90% right on Food-101, which shows how well attention processes can pick up on little differences.

Advantages:

- Better focus on distinguishing features
- Good performance on datasets with a lot of detail
- Less susceptibility to background noise

Limitations:

- More memory and processing power needed

2.4 Vision Transformers (ViTs)

Vision Transformers improve attention processes by using self-attention over visual patches instead of convolutions. They are great at modeling global contextual connections and do very well on massive datasets.

In food recognition:

- When trained on huge, varied datasets like Food-101, ViTs generally do better than CNNs.
- They pick up on long-range relationships and the overall structure of meals.

Limitations:

- Need big training datasets to avoid overfitting
- Costly to run
- Works well on medium-sized or small food datasets

2.5 State-Space Models (e.g., VMamba)

State-space architectures are the newest type of vision model. They strike a good compromise between modeling global dependencies and being efficient with computers.

Res-VMamba (Chen et al., 2024) showed that state-space models can do better than both CNNs and ViTs on fine-grained food datasets like **CNFOOD-241**.

Strengths:

- Good at modeling both local and global aspects
- Cheaper to run than Transformers
- Good at generalizing for fine-grained jobs

Limitations:

- Not very old; not much study has been done on it yet.

- There aren't many huge pretrained weights available.

2.6 Food Image Datasets

The properties of the dataset have a big effect on how well deep learning works. This thesis centers on three extensively utilized datasets exhibiting diverse levels of complexity:

Food-11

- 11 categories
- Low variation
- Commonly used for baseline comparisons

Food-41

- 41 mid-level classes
- Moderate fine-grained complexity
- Effective for evaluating model generalization

Food-101

- 101,000 images across 101 categories
- High intra-class variability
- Includes noisy labels
- Most widely used benchmark dataset

Other research (e.g., CAFSD) stress how hard it is to find many foods and how important it is to have huge, well-labeled datasets.

2.7 Quantitative Results Summary

Table 2.1 gives a summary of some of the most important research in the literature, including the kind of model, the dataset, the performance, the strengths, and the weaknesses.

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

2.8 Summary of Literature Review

The literature demonstrates ongoing advancements in food image recognition:

- CNNs are good and powerful starting points.
- CNNs with attention increase fine-grained categorization.

- Vision Transformers are better at modeling the world when there is a lot of data.
- State-space models find a good balance between performance and efficiency.
- The complexity of the dataset is very important for the accuracy of the model.

Even though there have been a lot of improvements, no current study offers a comprehensive comparison analysis of all major architecture families and various datasets with different levels of complexity, which is the main reason for this thesis.

Chapter 3

Related Work

This chapter looks at earlier research that has looked at how deep learning-based methods may be used to recognize food images. Chapter 2 gave a general idea of the designs, but this chapter

looks at how they were used in real life, the problems that came up, and the holes that led to the unified comparative analysis that this thesis suggests.

The literature review is divided into four main groups: CNN-based methods, attention-enhanced approaches, transformer-based models, and state-space architectures. The chapter ends with a look at the present problems and the necessity for a systematic comparison of many datasets.

3.1.1 CNN-Based Food Recognition Methods

Convolutional Neural Networks (CNNs) were a big part of the early research on recognizing food images. Deep Food (Liu et al., 2016) is a groundbreaking study that used GoogLeNet (Inception) on Food-101 and UEC-256. It did far better than hand-crafted feature methods and showed that learnt hierarchical representations are better.

Based on this work, Boyd et al. (2024) did a systematic evaluation of seven CNN architectures and found that DenseNet was the best when it was set up with the right hyperparameters. Their research emphasized the significance of architectural depth, feature reuse, and model stability in fine-grained food categorization.

Even while CNNs are good at providing robust baselines, previous research has shown that they have trouble telling the difference between categories with little visual variations. This shows that we need more expressive designs.

3.1.2 Attention-Enhanced CNN Approaches

Researchers came up with attention mechanisms to fix the problems with CNNs. These mechanisms let models focus on the parts of the image that provide them the greatest information. MAF-Net (Chen et al., 2025) is one of the most important models in this group. It uses multi-level feature fusion, self-attention refinement, KL-divergence optimization, and Subclass Center Loss.

MAF-Net performed at the highest level on Food-101, UEC-100, and Vireo-172, showing that attention processes work well for fine-grained categorization. More research backs up the idea that paying attention can help clear up misunderstanding between dishes that look identical by focusing on little details like texture, the form of the ingredients, and the patterns of the garnish.

But models that promote attention usually need more computing power and more datasets, which might make them less useful in real-time or embedded systems.

3.1.3 Transformer-Based Food Recognition

Vision Transformers (ViTs) changed the way things work by replacing convolutions with global self-attention across picture patches. ViTs have done quite well in classifying food, especially on big datasets like Food-101.

Even if they have some good points, the literature has pointed out certain problems with them:

- Need a lot of training data to avoid overfitting
- Cost more to run than CNNs
- Not always doing well on medium-sized datasets

ViTs are still a very important area of research in food identification, even though they are a strong alternative for modeling long-range relationships.

3.1.4 State-Space Models for Food Recognition

State-space models are the most recent type of deep learning architecture. Res-VMamba (Chen et al., 2024) combines residual learning with the VMamba backbone to make global-local feature extraction more efficient.

Res-VMamba did better than a few CNN and transformer baselines on the CNFOOD-241 dataset, showing that it has a lot of potential for detailed food analysis. However, it hasn't been widely adopted yet because it was only recently released and doesn't have many pretrained weights.

3.2 Multi-Food Detection and Dataset Challenges

Several research have investigated multi-food detection in addition to single-food categorization. Karabay et al. (2024) presented the CAFSD dataset and assessed YOLOv8-XL in the context of multi-food detection situations.

Their results show that multi-food photos are still more harder because of occlusion, overlapping objects, and messy backgrounds. These research further emphasize the significance of dataset quality and complexity in assessing model performance.

3.3 Limitations in Existing Research

A review of previous research identifies many gaps that underpin this thesis:

1. Most studies only look at one architecture family, which makes it hard to compare other models.
2. Comparisons between datasets are uncommon, even significant disparities in dataset complexity and noise.
3. There aren't many research that compare CNNs, attention-based models, transformers, and state-space models all at once.
4. The performance patterns as the dataset gets more complicated (Food-11 $\rightarrow$ Food-41 $\rightarrow$ Food-101) are still not well understood.

People frequently forget about practical things like the cost of computing, the capacity to grow, and the ability to apply to many other fields.

3.4 Summary

The literature shows that CNNs, attention-based architectures, transformers, and state-space models are making quick progress in recognizing food images. However, this research are still incomplete, with few comparisons across models and datasets.

This thesis fills up these gaps by doing a unified, systematic benchmark that tests cutting-edge models on three levels of dataset complexity.

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>