nutrients

MDPI

Article

An Explainable CNN and Vision Transformer-Based Approach for Real-Time Food Recognition

Kintoh Allen Nfor [1], Tagne Poupi Theodore Armand [2], Kenesbaeva Periyzat Ismaylovna [1], Moon-Il Joo [2] and Hee-Cheol Kim [1,2,*]

[1] Department of Computer Engineering, Inje University, Gimhae 50834, Republic of Korea; nforallen94@yahoo.com (K.A.N.); kenesbaevaperiyzat7@gmail.com (K.P.I.)
[2] Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea; poupiarmand2@gmail.com (T.P.T.A.); joomi@inje.ac.kr (M.-I.J.)
* Correspondence: heeki@inje.ac.kr; Tel.: +82-55-320-3720

Abstract: Background: Food image recognition, a crucial step in computational gastronomy, has diverse applications across nutritional platforms. Convolutional neural networks (CNNs) are widely used for this task due to their ability to capture hierarchical features. However, they struggle with long-range dependencies and global feature extraction, which are vital in distinguishing visually similar foods or images where the context of the whole dish is crucial, thus necessitating transformer architecture. Objectives: This research explores the capabilities of the CNNs and transformers to build a robust classification model that can handle both short- and long-range dependencies with global features to accurately classify food images and enhance food image recognition for better nutritional analysis. Methods: Our approach, which combines CNNs and Vision Transformers (ViTs), begins with the RestNet50 backbone model. This model is responsible for local feature extraction from the input image. The resulting feature map is then passed to the ViT encoder block, which handles further global feature extraction and classification using multi-head attention and fully connected layers with pre-trained weights. Results: Our experiments on five diverse datasets have confirmed a superior performance compared to the current state-of-the-art methods, and our combined dataset leveraging complementary features showed enhanced generalizability and robust performance in addressing global food diversity. We used explainable techniques like grad-CAM and LIME to understand how the models made their decisions, thereby enhancing the user's trust in the proposed system. This model has been integrated into a mobile application for food recognition and nutrition analysis, offering features like an intelligent diet-tracking system. Conclusion: This research paves the way for practical applications in personalized nutrition and healthcare, showcasing the extensive potential of AI in nutritional sciences across various dietary platforms.

Keywords: food image recognition; convolutional neural networks; vision transformers; explainability methods; personalized nutrition

Academic Editor: Christian Ritz

Received: 30 November 2024
Revised: 9 January 2025
Accepted: 16 January 2025
Published: 20 January 2025

Citation: Nfor, K.A.; Theodore Armand, T.P.; Ismaylovna, K.P.; Joo, M.-I.; Kim, H.-C. An Explainable CNN and Vision Transformer-Based Approach for Real-Time Food Recognition. *Nutrients* **2025**, *17*, 362. https://doi.org/10.3390/nu17020362

1. Introduction

Food recognition's emergence in computational gastronomy is an important area with potential applications in diet tracking, nutritional assessment, and food study [1]. Correctly identifying and classifying food has various uses, including portion control, nutritional analysis, Meal Frequency Log (MFLog), and other uses [2]. While there have been minor advancements in using traditional machine learning and deep learning in nutrition, some tasks, such as handling visually complex food presentations, improving

fine-grained classification, ensuring scalability for real-time applications, and addressing inconsistencies in food datasets, still require enhanced methods that can potentially improve the current state of the art in food recognition systems [3–5]. The current advancement in computer vision, especially the application of deep learning, particularly the CNNs, has enabled the creation of models that can recognize food-related items from images [6,7]. At the same time, architectures like ResNet, DenseNet, and Inception have delivered great feats in image recognition [8,9]. However, some challenges are associated with using CNN, such as its ability to detect only local features [10,11]. Thus, they have a problem obtaining a global context, which is required to distinguish between visually similar foods [11].

Vision Transformers (ViTs) have recently emerged as a promising alternative with the attention mechanisms to capture long-range dependencies in images [12]. With this capability, ViTs have shown higher performance in several computer vision tasks such as image classification, object detection, and segmentation [10–12]. Hybrid models, on the other hand, have the potential to meet and surpass the limitations of traditional approaches by merging the strengths of CNNs and transformers [13]. Many factors delay the accurate development of food recognition systems, one major factor being the complexity and inconsistency of food images [14,15]. Models find recognizing and classifying food items difficult because they can appear in different forms, presentations, and lighting conditions [16]. To further complicate this, there is also inconsistency in food recognition image sizes and quality across various datasets, which often causes biases and potentially twists the model's classification performance [17,18]. Efficient knowledge transfer is another significant challenge, even though pre-trained modes can significantly improve accuracy and generalization across diverse food recognition datasets [19–21]. However, food recognition tasks require considerable computational resources, memory, and skill to finetune these pre-trained models accurately, and scalability is also of great concern [22,23].

Besides preventing disease, nutrition has a significant role in treating and managing various health issues [6]. According to a report from the American Diabetes Association, 2020, medical nutrition therapy, which is integral to dealing with diseases like diabetes, is based on the principles of nutrition science [7]. Furthermore, the field is growing rapidly due to the combination of genomics and personalized nutrition. Personalized nutrition uses individual genetic information to recommend dietary habits [8]; this could significantly impact disease prevention and health improvement.

The field of nutrition has traditionally relied on observational studies and clinical trials, but the advent of AI has supplanted this. These technologies, including machine learning, deep learning, and data analysis, have the potential to uncover complex relationships in large datasets, identify patterns, and generate actionable information [9,10]. From dietary personalization to preventative models that predict disease, the potential uses of AI in nutrition are numerous and widespread. Integrating AI applications in nutrition facilitates the technological advancements reshaping the landscape of dietary interventions [11]. AI techniques hold immense promise in this data-driven era for revolutionizing how we understand, monitor, and optimize nutritional outcomes.

In nutrition, AI can be defined as applying conceptual algorithms, machine learning, and deep learning techniques to analyze, interpret, and make informed decisions from various datasets related to nutritional data, dietary patterns, and other health factors [11–13]. Leveraging machine learning algorithms, random forests can be used to analyze genetics and dietary data to understand how nutrients influence human genetic variations [14,15]. Moreover, collaborative filtering techniques are widely used in personalized nutrition recommendations, while deep learning methods such as convolutional neural networks (CNN) and transfer learning pre-trained models (Resnet, EfficientNet, . . .) assist in identi-

fying and classifying meals using food images to detect dietary patterns and, further, to assess nutritional content [16–18].

In this research, we adopted a hybrid model architecture to address these challenges by combining the strengths of CNNs and ViT. We use efficient knowledge transfer using pre-trained weights and advanced training techniques to enhance scalability and efficiency. Our hybrid model architecture uses ResNet50 [24] as the backbone with transformer encoders, combining CNN feature extraction capabilities with the transformer's attention mechanisms [25]. We aim to capture both local and global features in diverse food images, thereby improving recognition accuracy and robustness.

Pretrained weights provide an initial foundation for the model, which helps for effective knowledge transfer [1,19,20]. To adapt the model to the nuances of different food categories, we fine-tuned these pre-trained models on five different food recognition datasets [26]. We used mixed precision [27] and distributed training [28] to accelerate the training process and efficiently handle large datasets. Mixed precision training performs computations in half-precision while maintaining accuracy, reducing memory usage, and speeding up the training process [29–31]. Distributed training balances the model across multiple GPUs for efficiently handling large datasets [32]. At the same time, gradient accumulation allows us to use larger batch sizes even with limited GPU memory, improving training stability [32–34]. We evaluate our model on five diverse food recognition datasets. This research is needed to design diet monitoring systems that may be very useful in the fight against lifestyle diseases such as obesity, diabetes, hypertension, and many others [2,35]. Our work also brings a sense of hope for better diet improvement and, therefore, a more optimistic future for this technology. Besides protecting consumers in the food industry, food recognition is a useful tool for this goal [2,35–37].

We target food recognition to push the existing state-of-the-art datasets to be more consistent and improve fine-grained classification capability [38]. We additionally approach state-of-the-art food recognition by identifying the gaps in existing methods and providing a strong, effective, accurate alternative. We seek to establish new state-of-the-art food recognition technologies by fusing CNNs and ViTs, pre-trained weights, and advanced training techniques.

The primary purpose of this study is to develop a robust hybrid model that combines CNNs and Vision Transformers (ViTs) to address the challenges identified in food image recognition. Achieving this goal has led to several key contributions, which can be summarized as follows:

- Developing a hybrid CNNs and transformers-based model for efficient food recognition
- Enhance model accuracy and address the computational resources required for Vision Transformers using advanced training techniques such as mixed precision and distributed training.
- Exploring the capabilities of explainable AI techniques (Grad-CAM and LIME) to ease the understanding and interpretation of the machine learning models' decision-making process
- Integrating the developed model into a mobile application for real-time food monitoring, tracking, and dietary assessment tasks.

Our work advances computational gastronomy and nutrition analysis [2,35] to the wider community by tackling challenges associated with the variability of food images, dataset inconsistencies, and effective knowledge transfer. The solution provides a solid foundation for rapid and precise food image recognition, which may be used in dietary monitoring, dietary assessment of energy value in real-time, and food-related research [2,35–37].

This work is organized into sections, with related works in food recognition in Section 2. Section 3 presents a description of our proposed approach, and the experimental results and discussion are reported in Section 4. Section 5 is the conclusion and future work of the study.

2. Related Works

Food recognition has seen advancements with deep learning techniques, particularly CNNs and, more recently, Vision Transformers (ViTs). However, the increasing complexity and diversity of food items require finer-grained recognition [39,40], leading to the birth of many methods in this direction for higher accuracy and efficiency in recognition tasks.

2.1. CNN-Based Approaches

CNNs, with their ability to capture and extract hierarchical features from images, form the basis for early prevalent models in food recognition [41]. However, although effective, capturing long-range dependencies and global contextual relations within an image has been challenging for CNNs [42]. Tasks that require distinguishing between images with visual similarities, such as food recognition with complex dishes where ingredient interaction and context are crucial, make things more difficult for CNNs [41,42].

Rahmat et al. [43] applied transfer learning with AlexNet to adapt to pre-trained models that were used for recognizing specific cultural Malaysian food. However, their approach struggles with nuanced distinctions, which are needed for complete categories. It could not capture global contextual relationships even though they were able to reduce training time and use transfer learning. In the same light, Rahmat and Kutty [44] used the ResNet-50 architecture with deep residual learning to improve feature extraction even though they still faced challenges, as their approach still could not capture the intricate relationships between ingredients and the overall compassion of the dish.

Other methods, such as high-precision food classification variants of the Inception module [45], like DeepFood and the lightweight CNN [46], demonstrate the ongoing confidence in CNNs for food recognition. However, these approaches still have the limitations of the previous methods used. A multi-task network, developed by Feng et al. [47], called FoodNet could be a good approach. However, as the added complexity of multi-scale and label dependency learning can hinder performance in real-time or large-scale applications, they also illustrate scalability and efficiency challenges. Lastly, still built on CNN architectures like ResNet50, ResNet101, and SENet, Min, W. et al. [48] used the Progressive Region Enhancement Network (PRENet) with enhancements through progressive training and self-attention mechanisms, which was tested on the large-scale Food2K dataset. Prenet could handle fine-grained food recognition and showed strong performance across multiple benchmarks. Nonetheless, integration into resource-constrained systems may be further hampered by challenges in identifying items of diverse categories that appear visually similar and present high computational demands for training. Despite its few shortcomings, PRENet proves good scalability and generalization across various food-related tasks, making it an important contribution to food recognition research.

2.2. ViT-Based Approaches

The introduction of ViTs has significantly changed food recognition methods and offers a new approach to modeling relationships within an image. Unlike CNNs, which capture just local features, ViTs capture global dependencies and understand the broader context of an image. This makes them useful, particularly in distinguishing complex dishes and fine-grained details tasks.

Bianco et al. [49] introduced the use of ViTs for food recognition, and these models armed us with the capacity to overcome some of the challenges inherent in CNNs. However, while ViTs are powerful in modeling long-range interactions, their reliance on large amounts of data and computational resources can be a significant drawback, particularly in resource-constrained environments. Similarly, the approach taken in Min et al.'s ISIA Food-500 dataset study [50] used a stacked global-local attention network inspired by transformer architectures. While this method proves proficient at providing a balance between global and local features and detecting foods from different viewpoints, there might be some drawbacks inherent in characterizing fine-grained differences between visually similar food items. The study by Peng et al. [51] extends ViT-based methods, and was based on self-supervised learning (SSL). This study compares various SSL models, including DINO (DIstillation with NO labels ViT-based), against traditional and non-contrastive approaches. Their results highlight ViTs' ability to learn visual representations from unlabeled food images. Nevertheless, the models' applications are optimized for batch sizes and computational ability and requires longer training time. Presumably, ViTs have unique benefits, but these outcomes pose some questions regarding resource utilization concerns and generalization capabilities.

ViT-based models are a promising direction for food recognition because they can consider the entire image and the relations between the different segments. However, due to a significant requirement for computation power and shortcomings in capturing subtle differences, which may require complex methods to deal with, hybrid approaches are needed to integrate the strengths of both CNNs and transformers.

2.3. Hybrid Approaches

Hybrid models, which integrate CNNs with transformers or other advanced mechanisms, are promising advancements in food recognition that use both approaches' strengths.

The Multi-Task Guided Multi-View Attention Network (MVANet) developed by Liang et al. [52] takes a multi-task approach alongside multi-view attention for better identification of Chinese food dishes, their ingredients, and associated recipes. MVANet uses multi-view attention together with CNNs to gain a better view of the various features across tasks, as it delivers higher feature complexity compared to the traditional CNN. Nonetheless, this enhances the complexity of the model, making it less scalable or computationally tractable, especially in real-time domains or with large data sets. The second type of hybrid approach is the RES-VMAMBA, a model derived from residual learning and state space models. The second model is the hybrid model that was proposed by Chen et al. [53] and enhances fine, medium, and coarse-grain classification, having both local and global feature extraction techniques. Nevertheless, due to its intricateness, RES-VMAMBA may have the drawback of decreased extensibility and inability to encode transformers at their full capacity for modeling long input-output dependencies. Moreover, the joint learning model for the multi-task setting proposed by Liu et al. [54] enhances CNN-based learning with attention mechanisms to identify the food items and their ingredients simultaneously. On the one hand, the use of CNNs allows this approach to capture the food/ingredients pairing adequately; however, it may not fully leverage the global score and struggles with capturing dependencies, especially for the numerous ingredient dishes. A common approach to enhancing the utilization and modeling of the food recognition dataset is to adapt CNN and ViT models, and their limitations are evident in the Hybrid models. Nonetheless, it appears that the approaches pose several challenges concerning complexity, scalability, and completely capturing complex food content, meaning that more improvements are achievable in the future.

From the above-related studies, it can be observed that regardless of the employed approaches (CNN, ViT, or Hybrid), some limitations can still be identified in food recognition methodologies. We can clearly state that the traditional CNN-based food recognition processes fail to address long-range dependencies. In contrast, though computationally costly, ViTs consider global contexts and entail large-scale datasets. CNN/ViT combinations offer good results, but encoding the data in this way is problematic regarding scalability, complexity, and, most importantly, real-time response performance. These limitations require a solution that balances local and global feature extraction while optimizing computational efficiency, particularly for real-time mobile applications. Our proposed approach addresses these gaps by combining CNN and ViT strengths with optimization techniques for efficient, scalable food recognition.

3. Methodology

This section presents the approach to developing the proposed food image recognition model using an integrated CNN and vision-based model. The methodology is structured into several subsections, beginning with the data preparation process (datasets and preprocessing). Then, we discuss the proposed architecture detailing the chosen ViT and the CNN combined to form the proposed hybrid model. The training process is subsequently described, focusing on distributed and mixed precision techniques.

3.1. Data Preparation

Dataset Selection and Preprocessing

This study merged five datasets to create a dataset with diverse cultural backgrounds and cuisines for robust food image classification.

The Benchmark for Food Image Recognition (ETH Food101): The first benchmark in food image recognition was prepared with the ETH Food101 dataset [55]. It has 101,000 images and 101 food categories of Western and Asian cuisines, making it one of the most common databases in this area. One criterium of this dataset is its deliberately noisy labels in the training set. In real-world situations, data are frequently not well labeled; hence, this noise can mimic such situations, triggering our models to be robust. Additionally, all images are resized to 512 $\times$ 512 pixels. Given that Food101 has been widely used in research, it is considered a standard benchmark that paved the way for more specialized datasets to follow.

Cross-Cultural Food Recognition (VireoFood-172): Even though Food101 captured a range of common food items, the VireoFood-172 dataset [56] extends the collection to include other culturally relevant dishes, such as those in Asian cuisine. This dataset, containing 172 categories, tries to bridge the gap in current Western-centric datasets such as Food101, which lacks a range of diverse food items. These foods are generally eaten in Asian countries, so the VireoFood-172 dataset is appealing for cross-cultural food understanding. It attempts to push the limits of model generalization by increasing diversity in the food classes (including classes that are visually similar yet represent entirely different dishes). Therefore, VireoFood-172 becomes necessary to develop models suitable for a global setting.

Focusing on Chinese Cuisine (CNFOOD 241): The CNFOOD 241 [57] is a unique and specialized Chinese food recognition dataset with over 241 diverse food classes from Chinese cuisine. One of the special features of Chinese food is that it looks visually complex; a single dish can hold many ingredients, each adding to the image. However, with that increased complexity comes difficulty in food recognition for models since they must distinguish foods that can sometimes look similar or look like any other foods. CNFOOD 241 solves this problem by making available a concentrated dataset that re-

searchers can use to tune their models to identify Chinese dishes better. This uniqueness is useful to advance further food recognition targeting Chinese cuisine.

A Regional Perspective with Bounding Boxes (UEC-256): This dataset is unique because it offers a regional-specific viewpoint, as the images come from Japanese food sources. With more than 31,000 images and 256 categories, what is distinctive about this dataset is that it not only provides a list of food items but also the bounding box annotations for each food item within an image. This feature makes UE-256 [58] a classification and localization task that can be performed using the same data. The bounding box annotations available for this dataset are also key in modeling and detecting food items using the images. This is extremely useful when different food items are in a single image, as is the case with bento boxes that are common in Japanese cuisine.

Scaling Up Diversity (Food2k): The last food dataset, Food2k [48], extends the previous ideas to a larger, more varied set of food categories. Some details, like the exact architecture of what is covered under Food2k, can be somewhat varied, but in general, the emphasis is on covering a wider variety of food items and likely includes items more sparsely represented in other datasets. Food2k can be especially beneficial in examining the generalization capacity of models over an extensive diversity of common and rare dishes. Due to the diversity, models trained on Food2k are much better prepared to work in practice, where there can be numerous different kinds of food.

The above-presented datasets were merged into a single dataset to obtain a larger dataset used in this study, as shown in Table 1. This dataset provides a more comprehensive tool with solid cross-cultural grounds that fairly represents and generalizes the trained models across diverse populations, as indicated in Figure 1. From the initial datasets, we denote $X = \{x_1, x_2, \ldots, x_5\}$ the obtained set of images after merging and $Y = \{y_1, y_2 \ldots, y_N\}$ the corresponding categories. $x_1, x_2, \ldots, x_5$ contains diverse images grouped per category. Each dataset x_i, $i \in [1\text{–}5]$ contains different categories of food images y_i, $i \in [1\text{–}N]$ that are repetitive across the datasets. The merging process consists of creating a single set $Z = \{z_1, z_2, \ldots, z_N\}$ where each image set belongs to a unique category of image food and may result in a fusion of all images of the same category uniquely representing a class. The label set $Y = \{y_1, y_2 \ldots, y_N\}$ contains N food categories with no duplicates. For the merging process, necessary operations were considered to maintain the dataset's integrity and avoid bias or false information. All duplicate food categories were carefully analyzed before merging. A manual normalization was conducted on the food categories to obtain the final labels set. Some of the considerations are illustrated in Table 2 below:

Initially, 261 categories had more than one representation in the combined dataset. These categories were finally merged into 142 distinct classes. The obtained dataset resulted in 2651 distinct food categories. Figure 2 shows an overview of the image data contained in the dataset.

Table 1. Details of datasets used.

Dataset Name	No of Classes	Total Samples	Regions	Size (GB)	Ref.
Food2K	2000	1,036,564	Miscellaneous	64.2	[48]
Food101	101	101,000	Western	9.55	[55]
VireoFood172	172	110,241	Asia	1.52	[56]
CNFOOD-241	241	191,811	Chinese	9.94	[57]
UCE-FOOD 256	256	31,395	Japanese	3.97	[58]
Total	2770	1,471,011		89.18	

Table 2. Normalization of Food Categories Across Datasets.

SN	Normalized Category	Original Food Names	Datasets	Type of Operation
1	beef curry	(“Beef curry” vs. “beef curry”)	(‘VireoFood172’, ‘FOOD 256’)	Case Sensitivity
2	saozi noodles	(‘Saozi noodles’, ‘saozi noodles’)	(‘FOOD 2K’, ‘CNFOOD-241’)	Punctuation
3	rice	(‘Rice’, ‘Rice’, ‘rice’)	(‘VireoFood172’, ‘CNFOOD-241’, ‘FOOD 256’)	Spacing Issues
4	bibimbap	(‘bibimbap’)	(‘FOOD 256’, ‘FOOD 101’)	Uniform

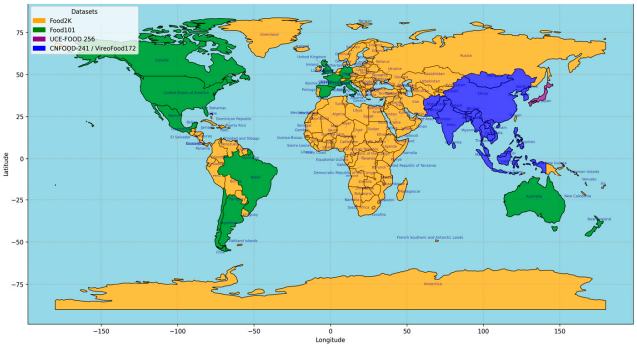

Figure 1. Global representation of food datasets.

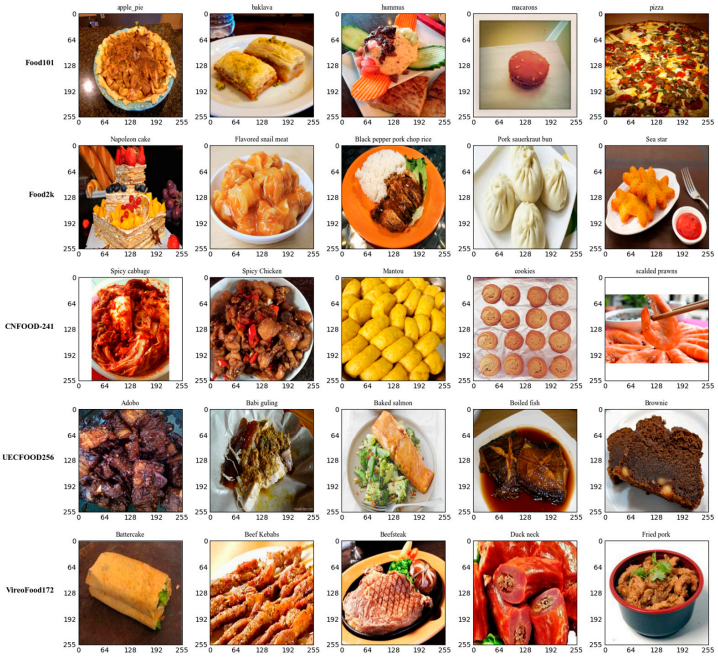

Figure 2. Samples from all the various datasets used.

Some preprocessing operations were carried out on the Z dataset. First, data were cleaned through duplicate removal, deletion of corrupt files (unreadable images), and filtering of irrelevant images (mislabeled, poor quality, or not food images). All the dataset images were resized to a standard size of 224 × 224 to match the input size requirements and ensure consistency in the input dimension for the model training. Furthermore, data normalization was conducted by pixel scaling, setting all pixel values to $[-1, 1]$ obtained by dividing all pixel values by 255. Additionally, data augmentation operations were implemented to improve model generalization and account for the diversity in the training set using random crop, horizontal flip, and color jittering.

3.2. Model Architecture

This research employs a hybrid model made of CNN and vision transformers.

3.2.1. Vision Transformer (ViT)

Vision transformers (ViTs) are a novel neural network derived from the transformer model adapted for image processing tasks. Though transformers were developed for natural language processing (NLP), the ViTs leverage a self-attention mechanism to analyze images and predict their class labels [59,60]. This is achieved by representing an input image as a series of image patches like the series of word embeddings used when applying transformers to text. Unlike the basic transformer, ViT uses only an encoder with a multi-layer perceptron for image classification. Figure 3 gives an overview that illustrates ViT, a key component of the hybrid model proposed in this study. This model includes patch embedding, a transformer encoder (made up of MHSA, residual connections, and FNN), and a classification head.

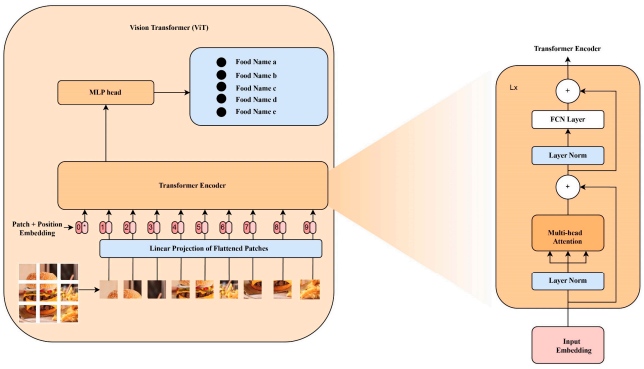

Figure 3. Vision Transformer (ViT) Architecture.

Patch Embedding

While CNNs apply filters across the entire image to extract important features, like edges, textures, or shapes, ViT conducts "patchifying", splitting the input image into smaller fixed-size partitions with no overlap. In the model, each patch is treated as a token, is flattened, and then projected to the output dimension, resulting in a sequence of patch embeddings. Positional information is added to each patch embedding to preserve encoded spatial data. For any input image x_i of size, $H \times W$, $\frac{H}{P} \times \frac{W}{P}$ patches of size P are generated and flattened into a 1D vector and projected into higher dimensional space through linear projection. The corresponding output embedding for the p^{th} patch after the linear transformation (Z_p) has a dimensionality D and is expressed in Equation (1):

$$Z_p = Flatten(x_i[p]).W_p + b_p, \quad (1)$$

W_p and b_p denote the weight matrix and the bias vector for the p^{th} patch.
The resulting patch embeddings are concatenated to form a matrix $Z \in R^{P \times D}$.

Transformer Encoder

The transformer encoder is the main vision transformer's block that sequentially processes images (herein food images), providing the global contextual understanding necessary for accurate classification. Due to the multitude of food styles, presentations, and backgrounds, coupled with feature scalability, the transformer is an ideal tool. It can effectively capture global dependencies via its self-attention mechanism and process image patches with spatial biases inherent in CNNS. Each transformer encoder block consists of the following:

- Multi-Head Self-Attention (MHSA)

MHSA is the main part of the transformer encoder. The attention mechanism creates interaction between a patch embedding and all other patches. Each patch generates Query (Q), Key (K), and Value (V) vectors used in the self-attention mechanism, allowing the model to weigh the importance of each patch compared to others, as shown in Figure 4. The self-attention mechanism helps ViT capture the image's global dependencies and distant relationships, thereby overcoming CNN challenges. The attention weights are computed using a scaled dot product expressed in Equation (2):

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

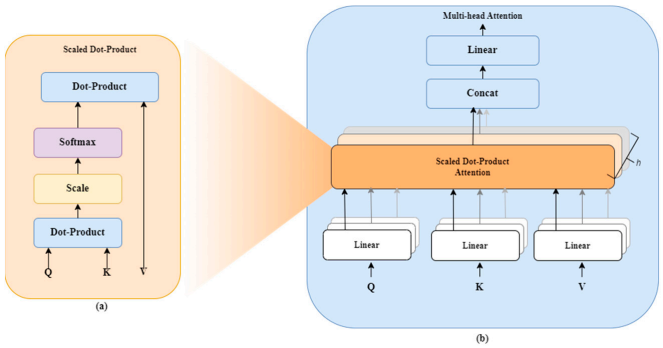

Figure 4. (**a**) Scaled dot-product attention and (**b**) multi-head attention, which consist of serval attention layers running in parallel.

With d_k denoting the dimension of the key vectors.

This self-attention mechanism provides a global view that allows the model to handle complex spatial interactions, such as those found in multi-object scenes or the fine patterns visible in food images.

Since there are different representation subspaces at every position, the multi-head attention layer aggregates the various attention heads by concatenation operations, as indicated in Equation (3).

$$\text{MultiHead}(\text{Q}, \text{K}, \text{V}) = \text{Concat}(\text{Head}_1, \ldots, \text{Head}_h)\text{W}_O \quad (3)$$

where h is the number of attention heads and W_O is the output weight matrix.

- Residual connections: This ensures stability during training and gradient flow. For a patch embedding X, the output is obtained using the following Equation (4):

$$X_{out} = X + MHSA(X) \quad (4)$$

- Feed-forward network (FNN): This is a two-layer Multi-Layer Perceptron (MLP) that uses the activation function to transform the output obtained from the attention mechanism. It is generalized and preceded by a layer normalization that ensures training stability. The FFN is computed as shown in Equation (5) below.

$$FFF(X) = \sigma(XW_1 + b_1)W_2 + b_2 \quad (5)$$

W and b are the trainable parameters.

Classification Head

After learning the image representations from the transformer encoder, the classification head is responsible for converting them into class predictions. In our case, this module is responsible for high-accuracy food recognition per category. Generally, the classification head contains a class token CLS representing the last transformer block and is used as a global feature vector. This token is passed through a fully connected layer to map the high-dimensional representation to the number of target classes. The SoftMax activation function generates the logits produced by the linear layer and then converts them into class probability, as in Equation (6).

$$P(c_i|X) = \frac{\exp(y_i)}{\sum_{j=1}^{C} \exp(y_j)} \quad (6)$$

where C is the number of food categories.

3.2.2. Transfer Learning

Transfer learning is crucial to this study as pre-trained models provide a strong initialization for the ViT, helping speed up convergence and improve generalization. These pre-trained models helped with a strong feature representation to be fine-tuned to our food classification task because they were trained on large datasets such as ImageNet.

Pre-Trained Models Used

In this study, the following pre-trained models were used:

- ViT-B_16: A fine-grained feature extraction of the 16 × 16 patch size ViT model, pre-trained on ImageNet, provides a robust initialization necessary for our task [59].
- ViT-B_32: This is another ViT model with a patch size 32 × 32, which provides an alternative between computational efficiency and fine-grained feature representation.
- R50-ViT-B_16: This research adopts the R50-ViT-B_16 architecture and integrates CNN and ViT to enhance image classification performance, specifically through a hybrid model combining ResNet-50 (R50) and Vision Transformer (ViT-B16).

We begin with a ResNet-50 backbone that serves as a feature extractor and has been modified by incorporating Squeeze-and-Excitation (SE) blocks, which recalibrate channel-wise feature responses to emphasize more informative features. It also includes an initial convolutional block followed by several stages of bottleneck residual blocks. After each bottleneck block, SE blocks are inserted to enhance the model's adaptive focus on important features. This backbone produces high-dimensional feature maps (2048 channels), crucial inputs to the Vision Transformer component.

Following the ResNet-50 feature extraction, our architecture includes a patch embedding layer to bridge the ResNet's output with the Vision Transformer's input requirements, which first reduces the channel dimensionality from 2048 to 1024 using a 1 × 1 convolutional layer. Then, a Conv2D layer with a stride equal to the patch size generates a sequence of patches from these feature maps, which are flattened to form a 2D sequence that serves as the tokens for the Vision Transformer.

A ViT-B16 encoder was used to capture an image's global information. It starts by prepending a learnable class token to the sequence of patch embeddings and includes positional embeddings so that spatial information among patches is not lost. A transformer encoder of twelve layers was used, each with multi-head self-attention that can simultaneously attend to different parts of the image. Each contains a feed-forward network, and the layers are interleaved with residual connections to increase non-linearity and stable training via layer normalization.

Lastly, the class token is passed through an MLP head for classification with global contextual enriched information. This head contains a densely connected linear layer that maps the class token to a specific number of output classes with a softmax activation giving out probabilities over these classes. The combination of these models (hybrid) has been shown to perform exceptionally well for tasks that need specific texture recognition and high-level image understanding because the ResNet-50 is good in localized feature extraction and the ViT complements it with global context.

These pre-trained models are chosen for their balance between computational efficiency and the ability to capture intricate patterns in image data. Figure 5 illustrates the adopted architecture as described above.

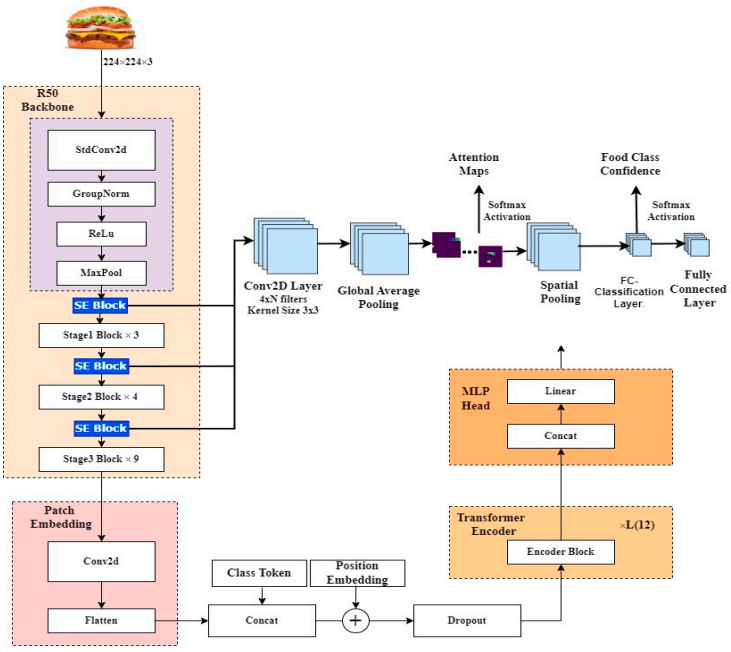

Figure 5. R50 + ViT-B_16 hybrid model.

Model Initialization with Pre-Trained Weights

To ensure the model can immediately start learning meaningful features without learning the basic structures from scratch, the weights for the patch embedding layer W_p and b_p are loaded from the pre-trained model, and positional embeddings E_{pos}, with initialized values from the pre-trained model, as shown in Equations (7) and (8). This is an important step because it permits our model to retain the special information about the patches essential for accurate image classification tasks.

$$W_p = LoadWeights(Model = ViT, Layer = PatchEmbeddings) \quad (7)$$

$$E_{pos} = LoadWeights(ViT, PositionEmbeddings) \quad (8)$$

The pre-trained weights are then used to initialize the transformer blocks, which include the multi-head attention layers and feed-forward networks, as shown in Equations (9) and (10). This entire process, as described, is visually illustrated in Figure 6, where the pre-trained model is loaded to ensure that the model does not learn features from scratch but immediately starts learning meaningful features using the source and target domain, with the MLP head used to classify the different food categories. This helps transfer the knowledge captured by the model during its pre-training on large-scale dataset weights or with weights fine-tuned on a similar classification task. In our case and for models pre-trained

on ImageNet, the classification head is adapted to the specific number of classes in each of the five datasets.

$$W_{MHSA} = \{LoadWeights\ from\ ViT,\ MultiHeadSelfAttention\} \tag{9}$$

$$W_{FFN} = \{LoadWeights\ from\ ViT,\ FeedForwardNetwork\} \tag{10}$$

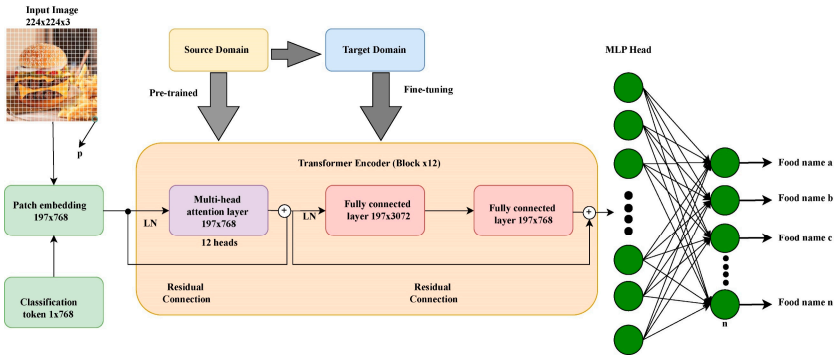

Figure 6. An architectural diagram of the pre-trained models used as a schematic representation of transfer learning.

Fine-Tuning Process

After implementing the pre-trained weights, the model is trained to classify food images. Because extreme updates may disrupt the pre-trained features, the learning rate is lower than the standard training from scratch. A typical choice is shown in Equations (11) and (12).

$$\eta_{fine\text{-}tune} = \frac{1}{10}\eta_{pre\text{-}train} \tag{11}$$

$$\eta_t = \eta_{fine\text{-}tune} \cdot Scheduler(t) \tag{12}$$

To retain their learned representations, the lower layers of the model, as shown in Equation (13), which capture more generic features, may be frozen at the initial stage and gradually throughout to allow the entire model to adapt to the new task; these layers are unfrozen during training. Regularization techniques such as dropout and weight decay, as in Equation (14), are used during fine-tuning to prevent overfitting as the model is fine-tuned on a relatively smaller dataset than its pre-training data.

$$Freeze(L_{lower}), Unfreeze\ after\ E_{warm\text{-}up} \tag{13}$$

$$L_{reg} = \lambda \sum_{i=1}^{P} ||W_i||^2 + Dropout(p) \tag{14}$$

This fine-tuning process helps the model to adjust its weights to the specific distribution of the datasets while using the rich feature representations learned during pre-training.

3.3. Training Procedure

For the training process, we set hyperparameters that will decide how our model can learn. The initial learning rate is set to $\eta_0 = 0.03$, and the learning rate scheduler adapts it during training by applying a weight decay term that regularizes the model, as illustrated in Equations (15) and (16).

$$\eta_t = \eta_0 \cdot Scheduler(t) \tag{15}$$

$$L_{reg} = \lambda \sum_{i=1}^{P} ||W_i||^2 \tag{16}$$

Second, the optimizer is stochastic gradient descent (SGD) with a momentum of μ, accumulating gradients over k steps before averaging and performing the update. We then clip the gradients not exceeding a maximum norm γ to avoid potential instability.

This is properly expressed in Equations (17)–(20).

$$v_{t+1} = \mu v_t - \eta \nabla_W L \tag{17}$$

$$W_{t+1} = W_t + v_{t+1} \tag{18}$$

$$\Delta W = \frac{1}{k} \sum_{i=1}^{k} \nabla_w L(x_i, y_i) \tag{19}$$

$$if \ \| \nabla WL \| > \gamma, \ \nabla WL = \frac{\gamma}{\| \nabla_W L \|} \nabla_W L \tag{20}$$

3.3.1. Distributed Training

Distributed training is applied to use multiple GPUs. The Distributed Data-Parallel (DDP) framework synchronizes the model parameters across GPUs, with the training loss computed as in Equation (21):

$$L_{total} = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i) \tag{21}$$

where n is the number of GPUs, and the gradients are averaged across all GPUs.

3.3.2. Mixed Precision Training

Mixed precision training is employed to enhance computational efficiency. The model is trained using a combination of 16-bit and 32-bit floating-point precision. The loss, as shown in Equation (22), is scaled by a factor s to maintain numerical stability, and the gradients are computed with scaled precision and then unscaled before updating the weights.

$$L_{scaled} = s \cdot L \tag{22}$$

3.3.3. Training Loop

Our training loop includes a batch size of 32 and an initial learning rate of 0.03 with SGD optimizer at a momentum of 0.9 and no weight decay. This is followed by a warmup strategy of 500 steps, transitioning into a cosine decay of over 100,000 total steps. The gradient accumulation was set to 1 step, and gradients were clipped to a maximum standard of 1.0. Mixed precision training with 16-bit floating-point (FP16) is supported via NVIDIA Apex at optimization level O2. For reproducibility, the training was initialized with a random seed of 42, and we saved the model checkpoints in the specified output directory.

4. Results and Discussion

This section will discuss how we set up our proposed approach's training and testing environments and results and compared its performance with other techniques.

4.1. Experimental Setup

Environment

For the experiment's environment, PyTorch in Spyder IDE was used, with APEX for mixed-precision training, allowing the model to use both 16-bit and 32-bit floating-point numbers on a 13th-gen Intel(R) Core (TM) i5-13400 2.50 GHz and 64 GB system RAM with a GeForce RTX 4080 GPUs of 16 GB RAM manufactured by Intel Corporation (Santa Clara, CA, USA). For distributed training, a system equipped with four NVIDIA

RTX A5000 GPUs manufactured by NVIDIA Corporation (Santa Clara, CA, USA), each with a computing capability of 8.6, supported by DDP, suggests that the environment can scale across multiple GPUs, possibly spread over multiple nodes.

4.2. Evaluation Metrics

Several metrics were used to assess the performance of our model:

- Accuracy: This metric measures the model's effectiveness by evaluating how many overall predictions were correct. The accuracy is computed as:

$$\text{Accuracy} = \left(\frac{TP + TN}{TP + FN + FP + TN}\right) \times 100 \tag{23}$$

- Recall: This measures the ability of a model to identify all relevant instances of a specific food class correctly, calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{24}$$

- Precision: That gives the proportion of images classified to a specific class that truly belongs to that class, provided by:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{25}$$

- F1 Score: This measures the test's accuracy, defined as the harmonic mean of precision and recall, computed as:

$$\text{F1 Score} = \left(\frac{2 \times Recall \times Precision}{Recall + Precision}\right) \times 100 \tag{26}$$

- Top-k Accuracy: A metric that represents how often the correct class is in one of the top-k predicted classes with the highest predictions:

$$\text{Top-k Accuracy} = \frac{1}{N} \sum_{i=1}^{N} 1\left(y^{(i)} \in P^{(i)}\right) \tag{27}$$

where N is the total number of samples in the dataset, $y^{(i)}$ is the true class label for the i-th sample and $P^{(i)}$ represents the set of top k predicted classes for the i-th sample.

4.3. Model Training, Testing, and Evaluation

To ensure a consistent evaluation during the training and testing phases of the pre-trained ViT models used in this study, a standardized set of initial parameters was used, with input image size set to 224 × 224 pixels, a standard ViT requirement. We performed a periodical assessment based on validation accuracy during the training, which was set to stop the training if no improvements were observed using the early stopping criterion based on the validation accuracy. Additionally, the validation dataset was used after a specified number of steps to evaluate the performance of our model. In optimization, a cross-entropy loss function was used, which is appropriate for the multi-classification characteristics of food recognition.

Figures 7 and 8 show the training and validation accuracy and top 5 accuracy curves for each pre-trained model and our proposed model on the VireoFood172 dataset. Figure 7 shows the hybrid model's steady improvement in training accuracy and superior validation accuracy, showcasing its ability to generalize effectively to unseen data. Furthermore, Figure 8 demonstrates the consistently higher top five accuracy achieved by the hybrid model across steps, reflecting its robustness in identifying the correct class within the

top five predictions, even for challenging food categories. We also plotted the F-1 Score, precision, and recall, as shown in Figures 9–11, respectively. The combined dataset's results were evaluated for Precision, Recall, and F1 Score, as shown in Figure 12. These plots provide details into the learning progress, showing how each model adapts to the different datasets used in this study. We divided the dataset into training and validation sets, maintaining an independent validation set to evaluate the model's performance and generalization, implementing hyperparameter optimizations, and avoiding over-fitting.

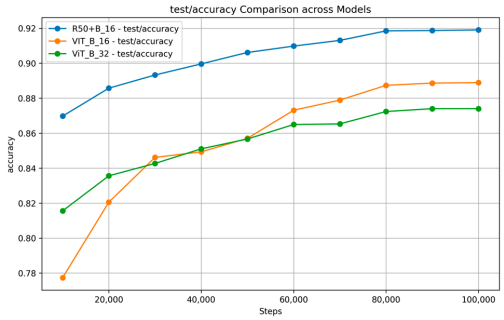

Figure 7. Accuracy Comparison across Models on VireoFood172.

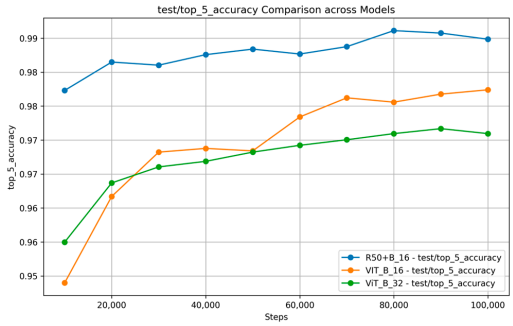

Figure 8. Top 5 accuracy Comparison across models on VireoFood172.

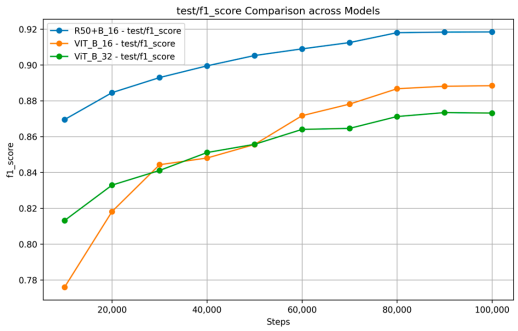

Figure 9. F-1 Score Comparison across Models on VireoFood172.

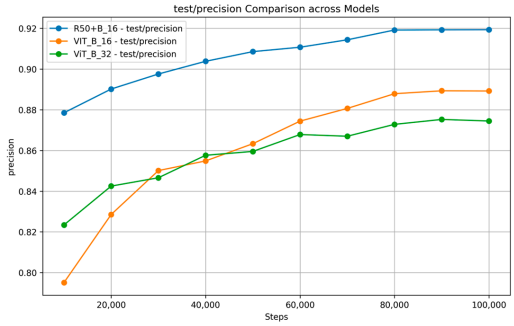

Figure 10. Precision Comparison across models on VireoFood172.

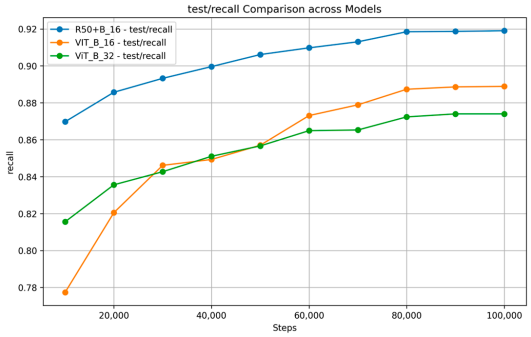

Figure 11. Recall Comparison across models on VireoFood172.

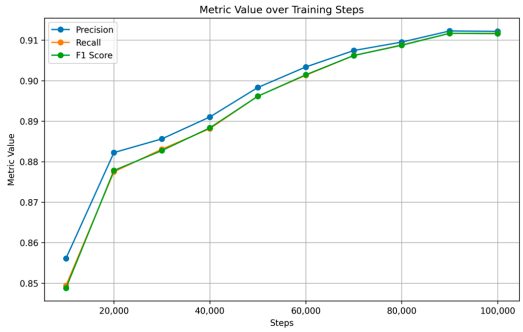

Figure 12. Precision, Recall, and F1 Score on Combined Dataset.

It should be noted that the R50 + ViT_B_16 model provides competitive results in terms of accuracy and top five accuracy in all datasets compared to ViT_B_16 and outperforms for these two metrics, whereas ViT_B_32 is always a worse player as shown in Tables 3 and 4. This demonstrates the hybrid's performance due to the powerful image representations that stand-alone ViT or CNN architectures could achieve, presumably on the back of the convolutional layers for feature extraction. Our combined dataset uses complementary features from individual datasets to achieve competitive performance and high generalizability, highlighting the importance of dataset fusion in developing robust models capable of addressing global food diversity and variability.

Table 3. Accuracy and Top five accuracy evaluation of our approach on six datasets.

Model	Food2k		Food101		VireoFood172		UCE-Food256		CBFOOD-241		Combined	
	Acc	T-5 Acc	Acc	T-5 Acc	Acc	T-5 Acc	Acc	T-5 Acc	Acc	T-5 Acc	Acc	T-5 Acc
VIT_B_32	68.3	91.1	84.3	96.2	88.0	97.2	81.0	95.2	69.0	91.3	---	
VIT_B_16	77.2	95.3	86.6	97.5	90.0	98.7	77.1	94.3	75.2	94.0	---	
R50 + VIT_B_16	84.1	96.2	91.3	99.0	92.3	98.5	85.0	98.0	83.4	95.2	91.17	98.35

Table 4. F1 Score, Recall, and Precision evaluation of our approach on six datasets.

Model	Food2k			Food101			VireoFood172			UCE-Food256			CBFOOD-241			Combined		
	F1 Score	Recall	Prec	F1 Score	Recall	Prec	F1 Score	Recall	Prec	F1 Score	Recall	Prec	F1 Score	Recall	Prec	F1 Score	Recall	Prec
VIT_B_32	63.4	64.1	63.3	86.0	86.0	86.0	89.1	89.1	89.0	94.9	94.0	94.9	76.0	76.0	76.7		-	
VIT_B_16	71.5	71.1	71.0	89.1	89.2	89.2	88.2	88.2	88.2	92.5	92.6	92.7	77.0	75.2	75.8		-	
R50 + VIT_B_16	84.2	84.1	84.0	91.1	91.2	91.3	93.7	93.8	93.8	95.3	95.5	95.3	83.4	83.5	83.4	88.35	91.17	91.22

As shown in Table 5, performance comparison across six datasets demonstrates competitive or superior accuracy compared to other works in the field. On the CBFOOD-241 dataset, we achieved an accuracy of 83.4%, which outperforms the results reported in [53]. Similarly, for the UCE-Food256 dataset, our model's accuracy of 85.0% is substantially higher than the accuracies reported by [44,45]. On the VireoFood172 dataset, we attained an accuracy of 93.7%, which exceeds the performance reported in [47,48,52,54]. This shows the effectiveness of our approach in handling complex food image datasets. In the case of Food101, our model reached 91.3%, showing superior results compared to [44–46], while maintaining competitive performance against [48,49]. For the Food2k dataset, our accuracy of 84.1% slightly surpasses the results in [48]. Our model performs robustly across all these diverse datasets, consistently outperforming or matching existing methods.

Table 5. Performance comparison across five datasets.

Dataset	Ref.	Technique	Accuracy (%)	Ours (%)
CBFOOD-241	[53]	RES-VMAMBA	82.15	83.4
UCE-Food256	[44]	ResNet-50	49.09	85.0
	[45]	Inception	80.7	
VireoFood172	[47]	FoodNet	89.73	93.7
	[48]	PRENet	90.80	
	[52]	MVANet	91.08	
	[54]	Attention Fusion Network (AFN)	89.54	
Food101	[44]	ResNet-50	39.75	91.3
	[45]	Inception	77.4	
	[46]	CNN	77.3	
	[48]	PRENet	91.13	
	[49]	ViT	92.59	
	[51]	Self-supervised	51.0	
Food2k	[46]	CNN	79.7	84.1
	[48]	PRENet	83.75	
	[49]	ViT	80.76	

4.4. Explainability

Explainability techniques help understand and interpret machine learning models' decision-making process, offering insights into which features or regions influence their predictions [61,62]. This work focuses on two popular methods: Grad-CAM and LIME.

4.5. Grad-CAM (Gradient-Weighted Class Activation Mapping)

In implementing Grad-CAM for the Vision Transformer (ViT)-based food classification model, hooks were used to capture the feature maps and gradients from a specific layer during forward and backward passes [61]. The gradients are averaged to generate weights, which are combined with the feature maps to create class activation maps (CAMs) highlighting the most important regions influencing the model's prediction, as shown in Figure 13. For ViT, the class token was excluded to focus only on image patches. The heatmap for a large input image is then resized to the dimensions of the corresponding input image and overlaid on it, such that we get an insight into which areas of a feature in the model played a major role during the decision-making by the model.

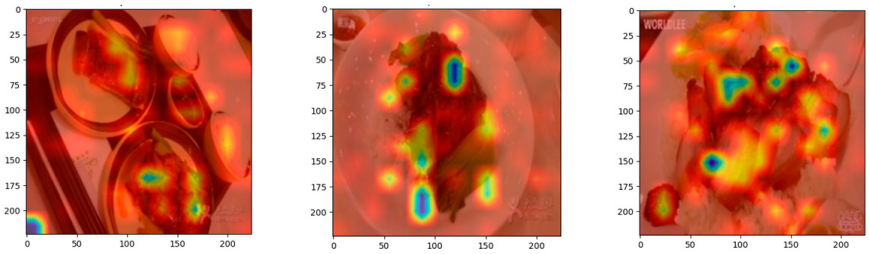

Figure 13. Visualization of Grad-CAM for explainability.

4.6. LIME (Local Interpretable Model-Agnostic Explanations)

LIME works by generating perturbations of the input image and observing the model's predictions for each perturbation, thus identifying which parts of the image contribute most to the prediction [63].

LIME generates perturbations of the input image and makes predictions about each perturbation [64]. LIME's image explainer was used to segment the image using random algorithms (SLIC, Felzenszwalb, or Quickshift), and the perturbations were applied to these segments, as shown in Figure 14. The model predicted class probabilities for each perturb image, which allowed LIME to understand how much each segment contributed to the final prediction. Afterward, we visualize the top contributing regions by overlaying a mask on the original image, highlighting the important areas for the predicted class.

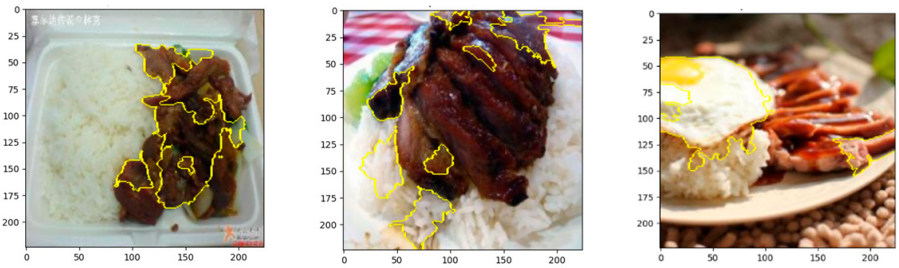

Figure 14. Visualization of LIME for explainability.

5. Model Integration

Subsequently, the trained model was converted into an ONNX (Open Neural Network Exchange), widely used for mobile deployment, to improve its performance on mobile devices. It was then integrated into a mobile application developed in React Native with Expo and Expo Router, as illustrated in Figure 15. The app enables users to take a photo of the food or choose a photo from their camera roll. The image captured or uploaded is passed through the model for recognition, which is initiated directly on the mobile application running the interface and does not require a constant internet connection to function. As shown in Figure 16, our trained model has been successfully integrated into a functional mobile application, demonstrating its ability for real-time dietary assessment.

Once the food is identified, the recognized food name is sent to the Edamam database, which contains information including, but not limited to, calories, macros, etc. The extracted nutritional information is then presented to the user, who can record the recognized food in the user's diary for monitoring. This helps make dietary information easy to follow and understand by the common user, resulting in easy tracking of nutritional habits for improved nutrition. Thus, the application is a comprehensive tool for food recognition and dietary tracking, enhancing the user experience through real-time feedback and accessibility.

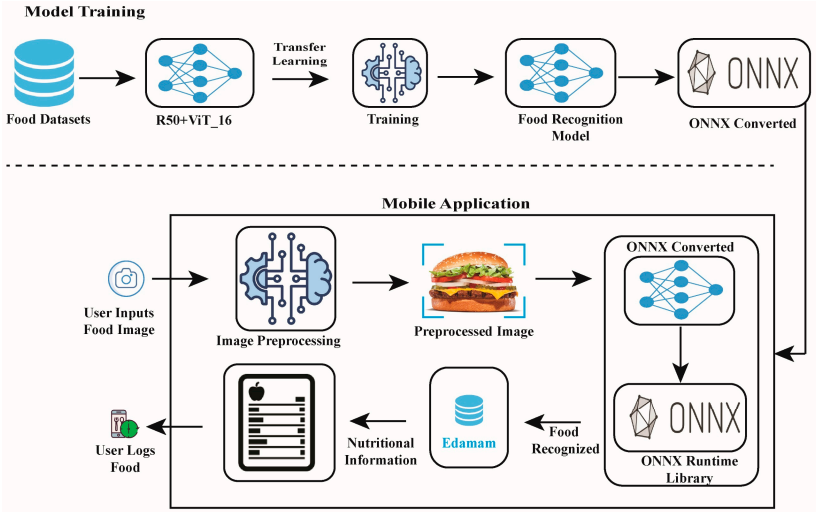

Figure 15. Mobile Application Integration Architecture.

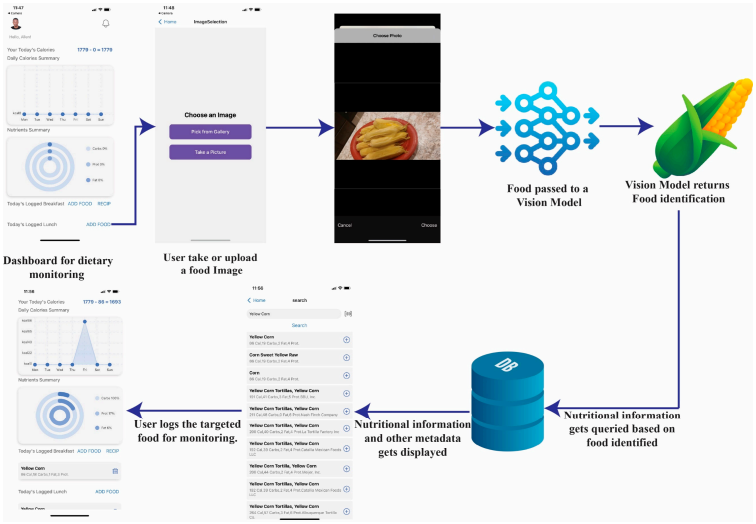

Figure 16. Integration into a mobile dietary monitoring, tracking, and nutritional assessment application.

6. Discussion and Conclusions

This study presented a hybrid food recognition model that integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to address critical challenges in food image classification. The hybrid model uses the strengths of CNNs in capturing localized features, such as edges and textures, and ViTs in modeling global dependencies and contextual relationships across the entire image. This combination allows the model to handle complex and diverse datasets with high variability in food presentation, ensuring more accurate and robust recognition. The evaluation conducted on five diverse datasets, including Food101, VireoFood172, CNFOOD-241, UCE-Food256, and Food2k, as well as a combined dataset, demonstrated the hybrid model's superior performance compared to state-of-the-art methods as shown in Table 4. The model consistently achieved higher accuracy and top five accuracy rates across all datasets. For instance, on the combined dataset, the hybrid model outperformed individual ViT-based and CNN-based models by effectively addressing the limitations of each approach, such as the inability of CNNs to capture long-range dependencies and the high computational cost of ViTs when used independently. The integration of Grad-CAM and LIME explainability techniques is a notable strength of this study. Grad-CAM generated class activation maps, visually highlighting the regions in the image that influenced the model's predictions. LIME complemented

this by segmenting the image into interpretable parts and assessing their contribution to the model's classification decisions. These techniques provided valuable insights into the model's decision-making process, enhancing transparency and user trust. Such explainability is essential for applications in personalized nutrition, where users and practitioners must understand the basis of dietary recommendations or assessments. The hybrid model was further validated through its integration into a mobile dietary monitoring and tracking application. This application directly enables real-time food recognition on mobile devices, where users can capture or upload food images processed by the model to identify food items and retrieve their nutritional information. This functionality has significant implications for personalized health management, allowing users to monitor their dietary habits, track nutritional intake, and make informed dietary choices conveniently. In addition to its technical contributions, this study addresses practical challenges in food recognition, such as variability in food presentations, diverse cultural contexts, and inconsistencies across datasets. The combined dataset, created by merging five benchmark datasets, provided a comprehensive and diverse training ground for the model. This approach ensured that the model could generalize effectively across different cuisines and food categories, addressing a major gap in existing food recognition research. This study establishes a new benchmark in food image recognition by integrating CNNs and ViTs into a hybrid framework. The model's enhanced performance, explainability, and practical implementation in a mobile application highlight its potential as a transformative tool in personalized nutrition and automated dietary systems. By addressing the limitations of existing methods and proposing a scalable, interpretable, and accurate solution, this study lays a strong foundation for future advancements in the intersection of artificial intelligence and food science. Despite its success, several challenges, such as the combined dataset, despite its diversity, may still have limitations in representing certain food categories or cuisines, which could influence the model's adaptability to unrepresented regions. Moreover, although the hybrid model exhibits scalability through distributed and mixed-precision training, further optimization is needed to enhance its deployment efficiency on resource-constrained devices. Future research should focus on linking food recognition with predictive modeling for health outcomes, such as obesity and diabetes, and integrating it with health monitoring tools for real-time dietary insights. Expanding dataset diversity and optimizing the model for edge computing will enhance accessibility and performance, paving the way for more versatile dietary assessment tools.

Author Contributions: Conceptualization K.A.N., T.P.T.A. and H.-C.K.; methodology, K.A.N., T.P.T.A. and M.-I.J.; validation, H.-C.K.; formal analysis, K.A.N., T.P.T.A., K.P.I. and H.-C.K.; writing—original draft preparation, K.A.N. and T.P.T.A.; writing—review and editing, K.A.N., T.P.T.A., K.P.I. and M.-I.J.; supervision, H.-C.K.; funding acquisition, H.-C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science ICT), Republic of Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information and Communications Technology Planning and Evaluation) in 2022 (2022-0-01091, 1711175863).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this research are publicly available, and the references are in the dataset description.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CNN	Convolutional Neural Network
ViT	Vision Transformer
Grad-CAM	Gradient-weighted Class Activation Mapping
LIME	Local Interpretable Model-agnostic Explanations
FP16	16-bit Floating Point
SGD	Stochastic Gradient Descent
DDP	Distributed Data-Parallel
SE	Squeeze-and-Excitation
ONNX	Open Neural Network Exchange
MFLog	Meal Frequency Log

References

1. Rokhva, S.; Teimourpour, B.; Soltani, A.H. Computer Vision in the Food Industry: Accurate, Real-time, and Automatic Food Recognition with Pretrained MobileNetV2. *arXiv* **2024**, arXiv:2405.11621. [CrossRef]
2. Armand, T.P.T.; Nfor, K.A.; Kim, J.-I.; Kim, H.-C. Applications of Artificial Intelligence, Machine Learning, and Deep Learning in Nutrition: A Systematic Review. *Nutrients* **2024**, *16*, 1073. [CrossRef]
3. Krutik, R.; Thacker, C.; Adhvaryu, R. Advancements in Food Recognition: A Comprehensive Review of Deep Learning-Based Automated Food Item Identification. In Proceedings of the 2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC), Setif, Algeria, 12–14 May 2024; pp. 1–6.
4. Mansouri, M.; Chaouni, S.B.; Andaloussi, S.J.; Ouchetto, O. Deep learning for food image recognition and nutrition analysis towards chronic diseases monitoring: A systematic review. *SN Comput. Sci.* **2023**, *4*, 513. [CrossRef]
5. Kiourt, C.; Pavlidis, G.; Markantonatou, S. Deep learning approaches in food recognition. In *Machine Learning Paradigms: Advances in Deep Learning-Based Technological Applications*; Learning and Analytics in Intelligent Systems; Springer: Cham, Switzerland, 2020; Volume 18, pp. 83–108.
6. Abiyev, R.; Adepoju, J. Automatic Food Recognition Using Deep Convolutional Neural Networks with Self-attention Mechanism. *Hum.-Centric Intell. Syst.* **2024**, *4*, 171–186. [CrossRef]
7. Fakhrou, A.; Kunhoth, J.; Al Maadeed, S. Smartphone-based food recognition system using multiple deep CNN models. *Multimed. Tools Appl.* **2021**, *80*, 33011–33032. [CrossRef]
8. Metwalli, A.-S.; Shen, W.; Wu, C.Q. Food image recognition based on densely connected convolutional neural networks. In Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 19–21 February 2020; pp. 027–032.
9. Sefer, M.; Arslan, B.; Batur, O.Z.; Sönmez, E.B. A comparative study of deep learning methods on food classification problem. In Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 15–17 October 2020; pp. 1–4.
10. Lin, X.; Lin, M.; Wei, L.; Chang, S.-F. Context-gated convolution. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XVIII 16; Springer International Publishing: Glasgow, UK, 2020; pp. 701–718.
11. Zahra, R.; Shekarizeh, S.; Sabokrou, M. Global-Local Processing in Convolutional Neural Networks. *arXiv* **2023**, arXiv:2306.08336.
12. Keller, M.; Tai, C.-e.A.; Chen, Y.; Xi, P.; Wong, A. NutritionVerse-Direct: Exploring Deep Neural Networks for Multitask Nutrition Prediction from Food Images. *arXiv* **2024**, arXiv:2405.07814.
13. Nijhawan, R.; Sinha, G.; Batra, A.; Kumar, M.; Sharma, H. VTnet+ Handcrafted based approach for food cuisines classification. *Multimed. Tools Appl.* **2024**, *83*, 10695–10715. [CrossRef]
14. Zhang, S.; Callaghan, V.; Che, Y. Image-based methods for dietary assessment: A survey. *J. Food Meas. Charact.* **2023**, *18*, 727–743. [CrossRef]
15. Aguilar, E.; Nagarajan, B.; Khatun, R.; Bolaños, M.; Radeva, P. Uncertainty modeling and deep learning applied to food image analysis. In *International Joint Conference on Biomedical Engineering Systems and Technologies*; Springer International Publishing: Cham, Germany, 2020; pp. 3–16.
16. Liu, G.; Yang, J.; Chen, J.; Zhu, B.; Jiang, Y.-G. From Canteen Food to Daily Meals: Generalizing Food Recognition to More Practical Scenarios. *IEEE Trans. Multimed.* **2024**, 1–10. [CrossRef]
17. Magomere, J.; Ishida, S.; Afonja, T.; Salama, A.; Kochin, D.; Yuehgoh, F.; Hamzaoui, I.; Sefala, R.; Alaagib, A.; Semenova, E.; et al. You are what you eat? Feeding foundation models a regionally diverse food dataset of World Wide Dishes. *arXiv* **2024**, arXiv:2406.09496.

18. Dalakleidi, K.V.; Papadelli, M.; Kapolos, I.; Papadimitriou, K. Applying image-based food-recognition systems on dietary assessment: A systematic review. *Adv. Nutr.* **2022**, *13*, 2590–2619. [CrossRef] [PubMed]
19. Singh, P.K.; Susan, S. Transfer Learning using Very Deep Pre-Trained Models for Food Image Classification. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; pp. 1–6.
20. Touijer, L.; Pastore, V.P.; Odone, F. Food Image Classification: The Benefit of In-Domain Transfer Learning. In *International Conference on Image Analysis and Processing*; Springer Nature: Cham, Switzerland, 2023; pp. 259–269.
21. Matarat, K. Enhancing Thai Food Classification: A CNN-Based Approach with Transfer Learning. *Math. Model. Eng. Probl.* **2024**, *11*, 1633–1640. [CrossRef]
22. Jiang, M. Food Image Classification with Convolutional Neural Networks. In *Deep Learning, Fall*; CS230; Stanford University: Stanford, CA, USA, 2019.
23. Al-Rubaye, D.; Serkan, A. Deep Transfer Learning and Data Augmentation for Food Image Classification. In Proceedings of the 2022 Iraqi International Conference on Communication and Information Technologies (IICCIT), Basrah, Iraq, 7–8 September 2022; pp. 125–130.
24. Kaiming, H.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Vaswani, A. Attention is all you need. In *Advances in Neural Information Processing Systems*; 2017.
26. Alshomrani, S.; Lina, A.; Banan, A.; Sarah, A.-S. Food Detection by Fine-Tuning Pre-trained Convolutional Neural Network Using Noisy Labels. *Int. J. Comput. Sci. Netw. Secur.* **2021**, *21*, 182–190.
27. Micikevicius, P.; Sharan, N.; Jonah, A.; Gregory, D.; Erich, E.; David, G.; Boris, G.; Michael, H.; Oleksii, K.; Ganesh, V.; et al. Mixed precision training. *arXiv* **2017**, arXiv:1710.03740.
28. Dash, S.; Isaac, L.; Yin, J.; Wang, X.; Egele, R.; Ellis, A.; Maiterth, M.; Cong, G.; Wang, F.; Balaprakash, P. Optimizing distributed training on frontier for large language models. In Proceedings of the ISC High Performance 2024 Research Paper Proceedings (39th International Conference), Hamburg, Germany, 12–16 May 2024; Prometeus GmbH: Lerici, Italy, 2024; pp. 1–11.
29. Lee, W.; Rahul, S.; Alex, A. Training with Mixed-Precision Floating-Point Assignments. *arXiv* **2023**, arXiv:2301.13464.
30. Marion, D.; Fan, M.; Andreas, M.K. Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks. *IEEE Access* **2023**, *11*, 57627–57634.
31. Aditya, R.; Vink, D.; Venieris, S.; Bouganis, C.-S. Multi-precision policy enforced training (MuPPET): A precision-switching strategy for quantised fixed-point training of CNNs. In Proceedings of the International Conference on Machine Learning, PMLR, Vienna, Austria, 12–18 July 2020; pp. 7943–7952.
32. Joel, L.-P. Layered gradient accumulation and modular pipeline parallelism: Fast and efficient training of large language models. *arXiv* **2021**, arXiv:2106.02679.
33. Zhang, Y.; Han, Y.; Cao, S.; Dai, G.; Miao, Y.; Cao, T.; Yang, F.; Xu, N. Adam Accumulation to Reduce Memory Footprints of both Activations and Gradients for Large-scale DNN Training. *arXiv* **2023**, arXiv:2305.19982.
34. Ma, Y.; Florin, R.; Wu, K.; Alexander, S. Adaptive elastic training for sparse deep learning on heterogeneous multi-gpu servers. *arXiv* **2021**, arXiv:2110.07029.
35. Theodore Armand, T.P.; Kim, H.-C.; Kim, J.-I. Digital Anti-Aging Healthcare: An Overview of the Applications of Digital Technologies in Diet Management. *J. Pers. Med.* **2024**, *14*, 254. [CrossRef]
36. Armand, T.P.T.; Deji-Oloruntoba, O.; Bhattacharjee, S.; Nfor, K.A.; Kim, H.-C. Optimizing longevity: Integrating Smart Nutrition and Digital Technologies for Personalized Anti-aging Healthcare. In Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan, 19–22 February 2024; pp. 243–248.
37. Tsolakidis, D.; Gymnopoulos, L.P.; Dimitropoulos, K. Artificial Intelligence and Machine Learning Technologies for Personalized Nutrition: A Review. *Informatics* **2024**, *11*, 62. [CrossRef]
38. Shao, Z.; He, J.; Yu, Y.-Y.; Lin, L.; Cowan, A.E.; Eicher-Miller, H.A.; Zhu, F. Towards the creation of a nutrition and food group based image database. *arXiv* **2022**, arXiv:2206.02086. [CrossRef]
39. Rodríguez-De-Vera, J.M.; Estepa, I.G.; Marc, B.; Bhalaji, N.; Petia, R. LOFI: LOng-tailed FIne-Grained Network for Food Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 3750–3760.
40. Boyd, L.; Nnamoko, N.; Lopes, R. Fine-Grained Food Image Recognition: A Study on Optimising Convolutional Neural Networks for Improved Performance. *J. Imaging* **2024**, *10*, 126. [CrossRef] [PubMed]
41. Khan, A.; Rauf, Z.; Sohail, A.; Khan, A.R.; Asif, H.; Asif, A.; Farooq, U. A survey of the vision transformers and their CNN-transformer based variants. *Artif. Intell. Rev.* **2023**, *56*, 2917–2970. [CrossRef]
42. Raghu, M.; Thomas, U.; Simon, K.; Zhang, C.; Alexey, D. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.

43. Rahmat, R.A.; Suhaili, B.K. Malaysian food recognition using alexnet CNN and transfer learning. In Proceedings of the 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, 3–4 April 2021; pp. 59–64.
44. Zharfan, Z.; Lee, C.P.; Lim, K.M. Food recognition with resnet-50. In Proceedings of the 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 26–27 September 2020; pp. 1–5.
45. Chang, L.; Cao, Y.; Luo, Y.; Chen, G.; Vokkarane, V.; Ma, Y. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *Inclusive Smart Cities and Digital Health: 14th International Conference on Smart Homes and Health Telematics, ICOST 2016, Wuhan, China, 25–27 May 2016*; Proceedings 14; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 37–48.
46. Li, X.; Li, Y.; Zou, X.; Ren, T. A high-precision food image classification method with a small number of parameters. In Proceedings of the 2023 4th International Conference on Computer, Big Data and Artificial Intelligence (ICCBD+ AI), Guiyang, China, 15–17 December 2023; pp. 33–36.
47. Feng, S.; Lu, Z.; Li, Y.; Han, C.; Gu, X.; Wei, S. Foodnet: Multi-scale and label dependency learning-based multi-task network for food and ingredient recognition. *Neural Comput. Appl.* **2024**, *36*, 4485–4501.
48. Min, W.; Wang, Z.; Liu, Y.; Luo, M.; Kang, L.; Wei, X.; Wei, X.; Jiang, S. Large scale visual food recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9932–9949. [CrossRef]
49. Bianco, S.; Marco, B.; Gaetano, C.; Paolo, N.; Flavio, P. Food Recognition with Visual Transformers. In Proceedings of the 2023 IEEE 13th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), Berlin, Germany, 3–5 September 2023; pp. 82–87.
50. Min, W.; Liu, L.; Wang, Z.; Luo, Z.; Wei, X.; Wei, X.; Jiang, S. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 393–401.
51. Peng, A.; He, J.; Zhu, F. Self-supervised visual representation learning on food images. *arXiv* **2023**, arXiv:2303.09046. [CrossRef]
52. Liang, H.; Wen, G.; Hu, Y.; Luo, M.; Yang, P.; Xu, Y. MVANet: Multi-task guided multi-view attention network for Chinese food recognition. *IEEE Trans. Multimed.* **2020**, *23*, 3551–3561. [CrossRef]
53. Chen, C.-S.; Chen, G.-Y.; Zhou, D.; Jiang, D.; Chen, D.-S. Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning. *arXiv* **2024**, arXiv:2402.15761.
54. Liu, C.; Liang, Y.; Xue, Y.; Qian, X.; Fu, J. Food and ingredient joint learning for fine-grained recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2480–2493. [CrossRef]
55. Lukas, B.; Guillaumin, M.; Gool, L.V. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings Part VI 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 446–461.
56. Chen, J.; Ngo, C.-W. Deep-based ingredient recognition for cooking recipe retrieval. In Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA, 15–19 October 2016; pp. 32–41.
57. Fan, B.; Li, W.; Dong, L.; Li, J.; Nie, Z. Automatic Chinese Food recognition based on a stacking fusion model. In Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 24–27 July 2023; pp. 1–4.
58. Kawano, Y.; Keiji, Y. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, 6-7, 12 September 2014*; Proceedings, Part III 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 3–17.
59. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
60. Armand, T.P.T.; Bhattacharjee, S.; Choi, H.-K.; Kim, H.-C. Transformers Effectiveness in Medical Image Segmentation: A Comparative Analysis of UNet-Based Architectures. In Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan, 19–22 February 2024; pp. 238–242.
61. Rudresh, D.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.* **2023**, *55*, 1–33.
62. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
63. Junior, K.J.; Carole, K.S.; Armand, T.P.T.; Kim, H.-C.; Initiative, T.A.D.N. Alzheimer's Multiclassification Using Explainable AI Techniques. *Appl. Sci.* **2024**, *14*, 8287. [CrossRef]
64. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should i trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.