# FoodViT-X: A Lightweight CNN–Transformer Hybrid for Food Image Classification

Mostafa Galib (22-49791-3), Meherunnesa Monami (22-49824-3), Tahmid Mahbub Tonmoy (22-49783-3) and Syed Mostafa Tahsinul Islam (22-49843-3)

American International University-Bangladesh, Dhaka, Bangladesh

## Abstract

Food image recognition is still a challenging thing becausse of visual variations, complex textures, and the issue of no fixed structural patterns in food images. Although convolutional neural networks (CNNs) have shown good performance by learning local visual features, they are limited in learning global contextual relationships between food components. In this work, we propose FoodViT-X, a lightweight hybrid CNN–Transformer model for food image classification. The model combines an Xception-based CNN backbone for local feature extraction with a shallow Transformer encoder that models global contextual information from convolutional feature maps. This design improves representation capability while keeping the architecture efficient and simple. FoodViT-X is evaluated on the Food-101 dataset using the official data split. The proposed model achieves 78.93% Top-1 accuracy, 94.19% Top-5 accuracy, 97.02% Top-10 accuracy, and a macro-F1 score of 78.99%, demonstrating balanced performance across food categories. These results indicate that lightweight CNN–Transformer hybrids can effectively enhance food recognition performance without relying on complex or computationally expensive architectures.

**Keywords:** Image Classification, Hybrid CNN-Transformer, Vision Transformers, Xception network, Deep Learning, Food Classification

## 1. Introduction

Food image classification aims to automatically recognize food categories from differenet visual data (images/videos). It has become an important task in applications such as dietary tracking, health monitoring, and food recommendation systems. Unlike many object recognition problems, food images often lack consistent shape or structure and show large variations in looks and shapes caused by ingredients, cooking methods, and presentation styles. These characteristics make food classification really challenging. CNNs have been widely used for food recognition because of their ability to learn discriminative local visual patterns such as texture and color. But, CNN-based models primarily focus on local information and fails to capture broader contextual relationships within an image, which are important for differentiating visually similar food categories. Transformer-based models address this limitation through attention mechanisms but typically require higher computational resources, limiting their practicality. This work explores a hybrid learning strategy that combines the strengths of both CNNs and Transformers. We propose FoodViT-X, a lightweight CNN–Transformer model that integrates an Xception backbone with a shallow Transformer encoder to enhance global feature reasoning while maintaining computational efficiency. The proposed approach aims to improve food classification performance without relying on complex or heavy architectures.

The main contributions of this work are as follows:

1) A lightweight hybrid CNN–Transformer model is introduced, where an Xception-based CNN extracts local food features and a shallow Transformer encoder models global contextual relationships from convolutional feature maps.
2) An efficient and well-evaluated food classification pipeline is developed and validated on the Food-101 dataset, demonstrating balanced performance.

## 2. Related Work

Food image classification has been widely studied due to its importance in dietary assessment and food analysis applications. A foundational contribution in this field is the Food-101 dataset introduced by Bossard *et al.* [1], which highlighted the challenges of food recognition arising from high intra-class variation, visually similar categories, and the absence of rigid spatial structure. Their work demonstrated that local color–texture components play a crucial role in distinguishing food types, although the proposed Random Forest–based framework relied on hand-crafted features and was later outperformed by CNN-based approaches.

With the advancement of deep learning, convolutional neural networks (CNNs) became the dominant solution for food image classification. Liu *et al.* proposed DeepFood [2], a CNN-based framework inspired by GoogLeNet that leveraged transfer learning to improve recognition performance across multiple food datasets, including Food-101 and UEC-256. While effective, DeepFood required high computational cost and benefited significantly from bounding-box annotations, limiting its practicality.

More recently, Gomes [3] introduced a transfer learning–based food classification approach using a fine-tuned Xception network on the Food-101 dataset. The model achieved strong classification performance through extensive data augmentation and efficient feature extraction. However, the approach remains limited to convolutional operations and does not explicitly model global contextual relationships within food images.

To address this limitation, Ma and Wang proposed DBRS-Net [4], a hybrid CNN–Transformer architecture that combines a ResNet branch for local feature extraction with a Swin Transformer branch for global contextual modeling. Although DBRS-Net demonstrates improved representation capability, it introduces higher computational complexity and is evaluated on a relatively small-scale dataset.

## 3. Methodology

### 3.1. Dataset

This study uses the Food-101 dataset, which consists of 101 food categories with 1,000 images per class, in toal 101,000 images. The dataset contains real-world food images captured under varying lighting conditions, backgrounds, viewpoints, and presentation styles. It is a challenging benchmark for food image classification.

We followed the official Food-101 train–test split. To enable model selection and prevent overfitting, the training set is further divided into training and validation subsets using a stratified split, ensuring class balance across splits. No external datasets or additional annotations are used in this work.

### 3.2. Data Processing

All images are resized to a fixed resolution of 224 × 224 pixels, which balances classification performance and computational efficiency. Each image is normalized to match the statistics of the pre-trained backbone network. During validation and testing, only resizing and normalization are applied to ensure fair and consistent evaluation. Images that cannot be properly decoded are safely skipped to avoid interruptions during training.

### 3.3. Proposed Architecture

The proposed FoodViT-X model is a lightweight hybrid CNN–Transformer architecture designed to capture both local texture features and global contextual relationships in food images.

**1) CNN Backbone (Local Feature Extraction)**

FoodViT-X employs an Xception network as its convolutional backbone. Xception uses depthwise separable convolutions, enabling efficient extraction of fine-grained texture and shape features while maintaining a relatively low computational cost. The backbone is initialized with ImageNet pre-trained weights, and only the final convolutional feature maps are retained for further processing.

**2) Feature Projection**

The output feature maps from the CNN backbone are projected to a lower-dimensional embedding space using a 1×1 convolution. This projection reduces channel dimensionality and prepares the features for transformer-based processing without altering spatial resolution.

**3) Transformer Encoder (Global Context Modeling)**

To model long-range dependencies and global contextual relationships, the projected feature maps are reshaped into a sequence of spatial tokens and passed through a shallow Transformer encoder. The Transformer consists of multiple self-attention heads and feed-forward layers, enabling the model to reason about relationships between different spatial regions of the image. Unlike transformer-heavy architectures, the Transformer in FoodViT-X operates on convolutional feature maps rather than raw image patches, allowing efficient global reasoning with minimal computational overhead.

**4) Classification Head**

The Transformer outputs are summed up using global average pooling across tokens, producing a compact global representation. This representation is then passed through a fully connected layer to predict the final food category.

## 4. Experimental Analysis

### 4.1 Experimental Setup

The proposed FoodViT-X model was implemented using the PyTorch framework. The Xception backbone was initialized with ImageNet pre-trained weights, while the Transformer encoder was trained from scratch. The model was optimized using the AdamW optimizer with a learning rate of $3 \times 10^{-4}$ and a weight decay of $1 \times 10^{-4}$. Cross-entropy loss was used as the objective function. Training was performed for 7 epochs with a batch size

of 32. A stratified split of the training set was used to create a validation subset for model selection. The model achieving the highest validation accuracy was saved and used for final evaluation on the test set. To provide a comprehensive performance assessment, multiple evaluation metrics were reported, including Top-1 accuracy, Top-5 accuracy, Top-10 accuracy, and macro-averaged F1-score. The macro-F1 metric was computed globally across the entire test set to ensure balanced evaluation across all food categories.

### 4.2 Results

Table I summarizes the performance of the proposed FoodViT-X model on the Food-101 test set.

**Table I: Performance of FoodViT-X on Food-101**

| Metric | Value |
|---|---|
| Top-1 Accuracy | 78.93 % |
| Macro-F1 Score | 78.99 % |
| Top-5 Accuracy | 94.19 % |
| Top-10 Accuracy | 97.02 % |

The results demonstrate that FoodViT-X achieves strong and balanced classification performance across all classes. The close alignment between Top-1 accuracy and macro-F1 score indicates that the model does not favor specific classes and generalizes well across the dataset. High Top-5 and Top-10 accuracies further suggest that the model effectively captures semantic similarities between visually related food categories. Compared to CNN-only approaches such as the Base Paper, which relies solely on convolutional feature extraction, the proposed hybrid architecture benefits from enhanced global contextual reasoning through the Transformer encoder. This improvement is achieved without introducing heavy transformer backbones or ensemble strategies, highlighting the efficiency of the proposed design.

Overall, the experimental results confirm that integrating lightweight Transformer-based reasoning on top of convolutional features can improve food image recognition performance while maintaining computational practicality.

## 5. Conclusion

This study presented FoodViT-X, a lightweight hybrid CNN–Transformer model for food image classification. By combining an Xception-based convolutional backbone with a shallow Transformer encoder, the proposed approach effectively captures both local texture information and global contextual relationships in food images while maintaining computational efficiency.

Experimental results on the Food-101 dataset demonstrate that FoodViT-X achieves balanced and reliable performance across all classes, as reflected by its strong Top-1 accuracy, macro-F1 score. The results indicate that incorporating lightweight attention-based reasoning can enhance food recognition performance beyond CNN-only architectures without introducing significant model complexity.

Future work will focus on evaluating the proposed approach on additional food datasets, exploring further efficiency optimizations, and investigating attention visualization techniques to improve model interpretability in real-world food recognition applications.

Code of the whole work can be accessible through this link : https://github.com/mostafa-galib/CVPR/tree/main/FINAL/Paper

| STUDENT ID | STUDENT NAME | CONTRIBUTIONS |
|---|---|---|
| 22-49791-3 | Mostafa Galib | Results, Methodology |
| 22-49824-3 | Meherunnesa Monami | Methodology, Conclusion |
| 22-49783-3 | Tahmid Mahbub Tonmoy | Introduction, Related Works |
| 22-49843-3 | Syed Mostafa Tahsinul Islam | Abstract, Introduction |

## References

[1] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – Mining Discriminative Components with Random Forests," in *Proc. European Conf. Computer Vision (ECCV)*, 2014, pp. 446–461.

[2] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment," in *Proc. Int. Conf. Smart Computing (SMARTCOMP)*, 2016, pp. 1–6.

[3] D. Gomes, "Classification of Food Objects Using Deep Convolutional Neural Network Using Transfer Learning," 2024.

[4] D. Ma and D. Wang, "DBRS-Net: A Hybrid Framework for Food Image Classification," 2025.