Intelligent Systems with Applications 29 (2026) 200632

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

A study on the generalization of DINOv2 features for food recognition tasks: A unified evaluation framework

Simone Bianco*, Marco Buzzelli, Gianluigi Ciocca, Flavio Piccoli, Raimondo Schettini

University of Milano-Bicocca, Italy

ARTICLE INFO

Keywords:
Semi-supervised learning
Food recognition
Food segmentation
Aesthetic assessment
Cross-domain adaptation
Continual learning

ABSTRACT

Self-supervised learning has recently gained increasing attention in computer vision, enabling the extraction of rich and general-purpose feature representations without requiring large annotated datasets. In this paper we aim to build a unified approach capable of deploying robust and effective analysis systems, replacing the need for multiple task-specific models trained end-to-end. Rather than introducing new architectures or training strategies, our goal is to systematically assess whether a single frozen self-supervised representation can support heterogeneous food-related tasks under realistic operating conditions. To this end, we performed an extensive analysis of DINOv2 features across multiple benchmark datasets and tasks, including food classification, segmentation, aesthetic assessment, and robustness to image distortions. In addition, we explore its capacity for continual learning by applying it to incremental food classification scenarios. Our findings reveal that DINOv2 features excel in many food-related applications. Their shared representations across tasks reduce the need for training separate models, while their strong generalization, high accuracy, and ability to handle complex multi-task scenarios make them a strong candidate for a unified food recognition approach. Specifically, DINOv2 features match or surpass state-of-the-art supervised methods in several food recognition tasks, while offering a simpler and more unified deployment strategy. Furthermore, they outperform end-to-end models in cross-dataset scenarios by up to +19.4% Top-1 accuracy and exhibits strong resilience to common image distortions by up to +48.0% robustness in Top-1 accuracy percentual difference, ensuring reliable performance in real-world applications. On average across all considered tasks, the DINOv2-based unified evaluation outperforms the state of the art by approximately 2.8% and 5.4%, depending on the chosen model size, while using only 6.2% and 23.9% of the total number of model parameters, respectively.

1. Introduction

The automatic recognition of food in images has become a cornerstone of many real-world applications, ranging from health monitoring and dietary analysis to food aesthetics assessment and waste reduction (Allegra et al., 2020; Liu et al., 2024; Wang, Zheng, et al., 2024). In these applications, tasks such as recognizing food items, segmenting their components, and assessing their properties (quantity, aesthetic quality, etc.) sometimes need to be carried out simultaneously. For instance, in compound plates, food recognition may need to segment, identify, quantify and describe the single components. These multiple processing stages are often to be carried out in resource-constrained devices like smartphones or embedded systems (Fakhrou et al., 2021; Kawano & Yanai, 2015; Kong et al., 2023) highlighting the need for approaches that are not only accurate but also efficient to train, update and deploy.

The automatic recognition of food images presents significant challenges that push the boundaries of computer vision research since food items have high intra-class variability (e.g., the same dish can appear differently across cultures), high inter-class similarity (e.g., visually similar but nutritionally distinct foods), and require accurate detection in cluttered backgrounds, that can be further affected by varying lighting conditions, and occlusions. Food recognition in real application scenarios requires therefore different tasks such as object localization, image segmentation, and image classification. These tasks are often carried out with traditional supervised learning approaches, such as Convolutional Neural Networks (CNNs) trained on large, annotated datasets (Chen et al., 2012; Min et al., 2020), achieving notable success in addressing, for example, single food recognition tasks. However, these methods face several limitations. First, the robustness of the methods depends on the use of large labeled data for training. The reliance

* Corresponding author.
E-mail addresses: simone.bianco@unimib.it (S. Bianco), marco.buzzelli@unimib.it (M. Buzzelli), gianluigi.ciocca@unimib.it (G. Ciocca), flavio.piccoli@unimib.it (F. Piccoli), raimondo.schettini@unimib.it (R. Schettini).

https://doi.org/10.1016/j.iswa.2026.200632
Received 31 October 2025; Received in revised form 15 January 2026; Accepted 2 February 2026
Available online 4 February 2026

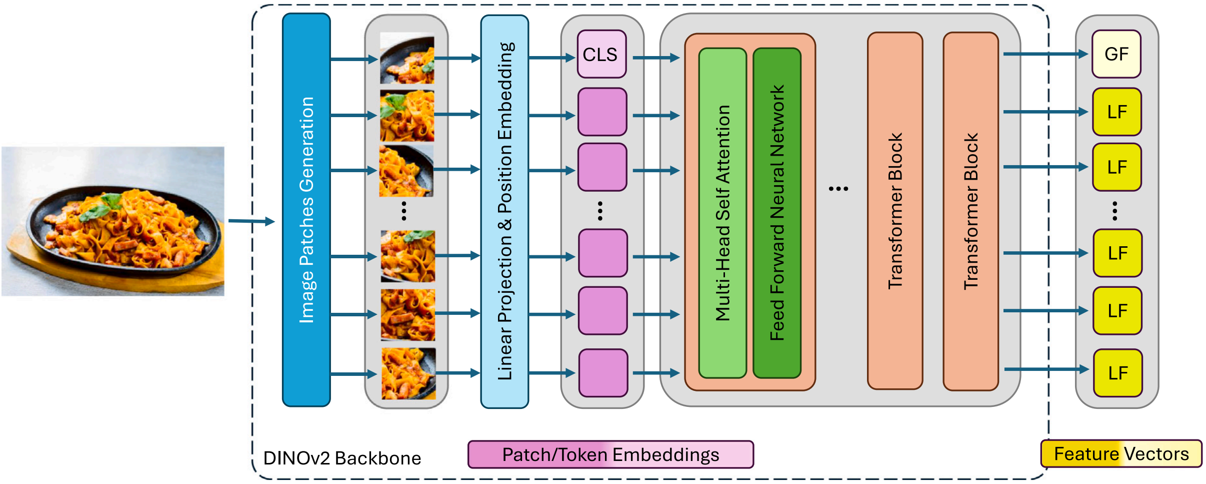


Fig. 1. Vision transformer architecture (Dosovitskiy, 2020) behind the DINOv2 model (Oquab et al., 2023). CLS: classification token. GF: global feature. LF: local feature.

on large-scale annotated datasets is resource-intensive and impractical for many applications. Moreover, traditional methods are often optimized for a single task, requiring separate models for segmentation, recognition, and food description, leading to increased computational overhead. Finally, extending these models to new categories or datasets corresponding to different food cuisines typically requires retraining or deep fine-tuning the model on new data, which can be computationally expensive and time-consuming.

Self-supervised learning (SSL) provides a powerful alternative to traditional supervised approaches, leveraging unlabeled data to learn rich feature representations and reducing the need for large manually labeled datasets (Gui et al., 2024; Liu et al., 2021). Several SSL frameworks have been developed in the last years, including SimCLR (Chen et al., 2020), MoCo (He et al., 2020), BYOL (Grill et al., 2020), and DINO (Caron et al., 2021). Among these, DINOv2 (Oquab et al., 2023) has emerged as a particularly effective solution, offering robust and versatile features for various image recognition tasks. DINOv2 is a self-supervised vision model built on a Vision Transformer (ViT) backbone (Dosovitskiy, 2020) that balances computational efficiency and task performance, making it well-suited for applications requiring both accuracy and adaptability. An input image is first resized and partitioned into non-overlapping patches, each of which is projected into a latent embedding space. These patch embeddings, together with a learnable "[CLS]" classification token and positional encodings, are processed by multiple transformer layers composed of self-attention and feed-forward networks. The self-attention mechanism allows the model to aggregate information across the entire image, producing a global representation via the [CLS] token as well as spatially localized patch-level features. Fig. 1 illustrates the architecture of the DINOv2 model and how its embeddings can be used to perform different tasks.

Throughout this work, the DINOv2 backbone is kept frozen, and the extracted features are used as input to lightweight task-specific heads for downstream food-related tasks: global features are to be used in image-level tasks (e.g., classification), since they provide a description of the whole image, while local features are to be used in patch-level tasks (e.g., segmentation), since they provide rich spatial information. This capability makes DINOv2 a versatile solution for scenarios requiring high-level semantic understanding and spatial delineation. While the strong out-of-the-box performance of DINOv2 has already been demonstrated on a variety of vision benchmarks, its systematic evaluation across multiple food-related tasks and learning regimes remains largely unexplored.

In this paper, rather than focusing on the design of new task-specific architectures, we investigate the extent to which a single large self-supervised visual representation can generalize across heterogeneous food recognition tasks. To this end, we extensively evaluate DINOv2 on a range of food recognition tasks with different complexities that reflect real-world food-related applications, highlighting its effectiveness compared to traditional task-specific models. Specifically, the tasks we considered are the following.

Food Segmentation. The food/no-food segmentation task (i.e., the task of classifying each image pixel as belonging either to food or to non-food regions, thus enabling automatic isolation of food items from the background) and the semantic segmentation task (i.e., the fine-grained task where each pixel is assigned to a specific food category like pasta, salad, meat, etc., thus enabling detailed food recognition and nutritional analysis) require precise pixel labeling, often necessary for localization and recognition of food, and for food quantity estimation. These tasks are tested using datasets `Food-50` (Chen et al., 2012) and `FoodSeg103` (Wu et al., 2021).

Food Classification. Food classification is a fundamental task in different applications and consists in automatically recognizing and assigning a food image to a predefined category (e.g., pizza, sushi, salad). This task is tested both on single datasets and in a cross-dataset scenario using `Food-50` (Chen et al., 2012) and `ISIA-Food500` (Min et al., 2020) datasets.

Food Aesthetic assessment. The aesthetic evaluation of food, i.e., automatically evaluating the visual appeal of food images using computational models, is critical in domains like social media and food marketing. This task is tested using the `Gourmet Photography Dataset` (Sheng et al., 2018).

Continual Learning. Real-world systems must often adapt to new categories without retraining on an entire dataset. This task is tested using the `ISIA Food-500` (Min et al., 2020) dataset.

Food recognition under distortions. Practical applications involve images with distortions such as noise, blur, or illumination changes. DINOv2's robustness is analyzed under challenging conditions using the `ISIA Food-500` (Min et al., 2020) dataset.

Fig. 2 shows how the features extracted from the DINOv2 backbone are used in the different food recognition tasks, as well as the achieved performance compared to methods in the state-of-the-art. It is important to underline that for DINOv2 the features are extracted from the same pre-trained model (i.e. DINOv2-B), while for the previous state of the art performance, a different method is considered for each task. We chose DINOv2-B since it is the best performing on average on the tasks considered, although other variants can obtain better results on individual tasks.

In summary, the contributions of this paper are as follows:

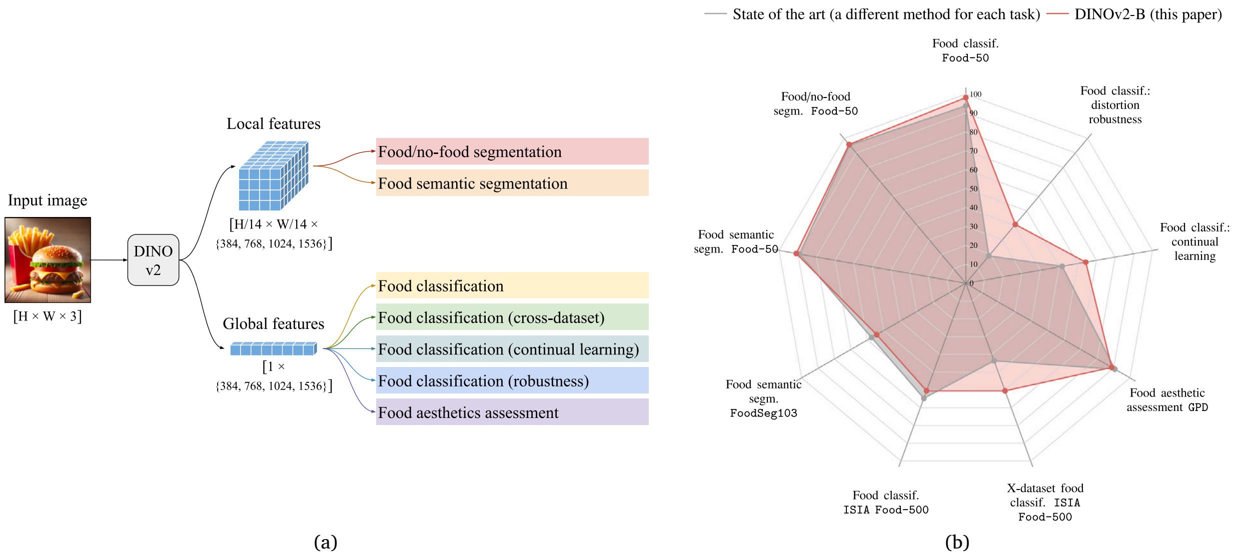

Input image
[H × W × 3]
DINO v2
Local features
[H/14 × W/14 × {384, 768, 1024, 1536}]
Global features
[1 × {384, 768, 1024, 1536}]
Food/no-food segmentation
Food semantic segmentation
Food classification
Food classification (cross-dataset)
Food classification (continual learning)
Food classification (robustness)
Food aesthetics assessment
(a)
State of the art (a different method for each task)
DINOv2-B (this paper)
Food classif. Food-50
Food classif.: distortion robustness
Food classif.: continual learning
Food aesthetic assessment GPD
X-dataset food classif. ISIA Food-500
Food classif. ISIA Food-500
Food semantic segm. FoodSeg103
Food semantic segm. Food-50
Food/no-food segm. Food-50
0
10
20
30
40
50
60
70
80
90
100
(b)

Fig. 2. (a) Depiction of how the features extracted from the DINOv2 backbone can be used in a unified approach for different food recognition tasks. H and W denote the width and height of the image, respectively. (b) Performance comparison between the DINOv2-B features and previous state-of-the-art methods for the various food-related tasks. Please note that for DINOv2 the features are extracted from the same pre-trained model (i.e. DINOv2-B), while for the previous state of the art performance, the best existing method is considered for each task. The average performance of state of the art methods is 67.24%, while that of DINOv2-B features is 72.66% (i.e., 5.42% higher).

- A systematic empirical analysis of the unified usage of pre-trained DINOv2 self-supervised features to address different food recognition tasks, demonstrating how pre-trained self-supervised features can serve as a common backbone across multiple tasks without requiring task-specific model architectures.
- Extensive experiments across several benchmark datasets showing that DINOv2 features achieve competitive or superior performance compared to traditional supervised models, particularly in cross-dataset scenarios.
- A comprehensive assessment of the DINOv2 features in a continual learning scenario showing significative performance improvement, especially in the few-shot regime.
- A thorough analysis of the DINOv2 features against different image distortions showing the robustness of the features in maintaining higher accuracy with respect to the other approaches, both in the case of single and multiple distortions.

The paper is organized as follows. Section 2 provides an overview of the state of the art for the considered food recognition tasks. The experimental setup is described in Section 3, while the experimental results are reported in Section 4. Finally, Section 5 discusses the practical implications for real-world food recognition and Section 6 draws the final conclusions.

2. Related works

Existing work on food recognition has achieved strong performance by designing task-specific models trained in a supervised manner for individual problems such as classification, segmentation, or aesthetic assessment. However, these approaches typically require large annotated datasets, do not generalize well across datasets, and must be retrained or fine-tuned when tasks or domains change. Moreover, the literature largely treats these tasks in isolation, with limited exploration of whether a single representation can support multiple food-related tasks effectively. In this work, we investigate whether self-supervised DINOv2 features can address these limitations by providing a unified, general-purpose representation applicable across tasks, datasets, and learning scenarios. In the following subsections we give a survey of the most relevant works in the literature about the food recognition tasks considered in this paper.

2.1. Food segmentation

Food segmentation, a key task in automatic food recognition, aims to delineate food regions in images, providing essential information for subsequent recognition and portion estimation. Food image segmentation has garnered significant attention in recent years due to its applications in dietary assessment, nutritional monitoring, and food recognition systems. Early work by Dehais et al. (2016) focused on segmenting and recognizing multiple food items in meal images to facilitate carbohydrate counting, highlighting the challenges in distinguishing overlapping and mixed food items. Aslan et al. (2020) conducted a comprehensive benchmark of various deep learning-based food segmentation methods using their proposed Food50Seg dataset. Both food/no-food segmentation and semantic segmentation approaches were evaluated. Evaluation was conducted from many different perspectives highlighting pros and cons of the methods.

To address the need for comprehensive datasets, Ege et al. (2019) proposed a new large-scale food image segmentation dataset aimed at improving food calorie estimation, emphasizing the importance of detailed segmentation in nutritional assessments. Also Wu et al. (2021) introduced a large-scale benchmark for food image segmentation, providing a foundation for training and evaluating segmentation models. Okamoto and Yanai (2021) release a dataset of 10,000 images annotated with complete segmentation masks and bounding boxes for training and testing state-of-the-art semantic segmentation models.

Recent advancements have explored the integration of transformer-based architectures with convolutional neural networks. Sinha et al. (2023) investigated transferring knowledge for food image segmentation using transformers and convolutions, achieving notable improvements in segmentation performance. Similarly, Lan et al. (2023) proposed FoodSAM, a framework that combines semantic masks with the Segment Anything Model by Kirillov et al. (2023) to enhance segmentation quality, demonstrating the potential of promptable segmentation in the food domain.

To tackle the problem of obtaining pixel-level annotations, Vlachopoulou et al. (2023) presented a weakly supervised methodology for food image classification and segmentation. Their approach utilizes attention-based multiple instance learning to generate semantic heat maps, reducing the reliance on detailed annotations. Due to the costly problem in training a semantic segmentation model because it requires a large number of images with pixel-level annotations (Honbu & Yanai, 2022) propose an unseen class segmentation method with high accuracy by using both zero-shot and few-shot segmentation methods for any unseen classes.

In the realm of real-time applications, Nguyen et al. (2024) developed FoodMask, a system capable of real-time food instance counting, segmentation, and recognition, highlighting the feasibility of deploying segmentation models in practical settings.

2.2. *Food recognition*

Food recognition is a fundamental aspect of visual analysis, essential for obtaining accurate food item labels in dietary assessments. Traditional methods often struggled with the inherent variability in food images, leading to low classification accuracy. In contrast, modern approaches utilize deep learning techniques, enabling end-to-end learning where neural networks automatically extract relevant features and perform classification tasks on food images.

One of the first work in food recognition that utilizes CNNs is (Kagaya et al., 2014). Through parameter optimization, the CNN demonstrated significantly higher accuracy in both food detection and recognition tasks compared to traditional support vector machine (SVM) methods utilizing handcrafted features. Notably, analysis of the convolutional kernels revealed a predominance of color features in the extraction process, underscoring the importance of color in food image recognition. Mezgec and Koroušić Seljak (2017) developed NutriNet, a deep convolutional neural network architecture tailored for food and drink image detection and recognition. The network is trained on a dataset of 225,953 images spanning 520 distinct food and drink categories. Bianco et al. (2023) trained and fine-tuned different vision transformer architectures on Food2K, a large-scale dataset of food images with 2000 categories, and compared the performance of vision transformers with convolutional neural networks on Food2K and Food101.

In Ciocca et al. (2017) the problem of food recognition in a canteen scenario is addressed. The food is put on trays and users select their meal within a set of daily dishes. The food regions are detected using traditional segmentation algorithms, while food recognition is performed locally the detected regions using a set of hand-crafted and CNN-based features coupled either with a k-NN or SVM classifier.

Food images pose peculiar challenges with respect to other image domain since they do not exhibit distinctive spatial arrangement and common semantic patterns. To address this problem, Jiang et al. (2019) designed a recognition approach that integrates high-level semantic features, mid-level attribute features, and deep visual features into a unified representation, capturing the semantics of food images across different granularities. The approach utilizes ingredient-supervised CNNs to extract mid-level attributes based on ingredient information, while class-supervised CNNs derive high-level semantic and deep visual features.

Enhancing food recognition using ingredients and other ancillary information about food has been widely studied. Marın et al. (2021) created Recipe1M+, a large-scale dataset containing over one million cooking recipes and 13 million food images, making it one of the largest publicly available collection of its kind. Using this dataset, a neural network was trained to create a joint embedding of recipes and images, achieving strong performance in image-recipe retrieval tasks. Adding a high-level classification objective further improved retrieval accuracy, rivaling human performance and enabling semantic vector arithmetic. Liu et al. (2020) proposed an fusion approach to generates the feature embeddings jointly aware of the ingredients and food. To this end, an Attention Fusion Network (AFN) and a Food-Ingredient Joint Learning module were developed. AFN identifies key food regions, while the joint learning module addresses ingredient imbalance using balanced focal loss. The method takes full advantage of multi-label ingredients information and improves the learning ability of the model. Shukor et al. (2022) proposed T-Food framework introducing a novel regularization scheme that exploits modality interaction while using unimodal encoders for efficient retrieval. It includes a dedicated recipe encoder to capture intra-dependencies and new triplet loss variants with dynamic margins for task difficulty adaptation. Leveraging Vision and Language Pretraining (VLP) models like CLIP for image encoding, T-Food significantly outperforms existing methods on the Recipe1M dataset.

Due to the very large food categories that exist, having a large dataset of food images is mandatory. Examples of this are the following. Ciocca et al. (2018a) introduced Food-475, the largest publicly available food dataset at the time, comprising 475 food classes and 247,636 images from four merged databases. Using a ResNet-50-based CNN, features were extracted and evaluated for food classification and retrieval. Results showed that features from Food-475 outperform others, highlighting the need for larger, more representative datasets to improve food recognition performance. The ISIA Food-500 dataset by Min et al. (2020) contains 500 categories and 399,726 images, surpassing existing benchmarks in size and diversity. A stacked global–local attention network was proposed, combining global-level features (texture, shape) with local-level features (ingredient regions) using hybrid attention and spatial transformers. Experiments on ISIA Food-500 demonstrated the model’s effectiveness, establishing it as a strong baseline for food recognition. Min et al. (2023) introduced Food2K, the largest food recognition dataset with 2000 categories and over 1 million images, setting a new benchmark for food visual representation. A deep progressive region enhancement network is proposed, featuring progressive local feature learning and region feature enhancement through self-attention. Extensive experiments show the effectiveness of this method, with Food2K demonstrating strong generalization across various food-related tasks and offering potential for future applications in food nutrition and other complex tasks.

Recent work continues to propose new benchmarks and methods, including Res-VMamba (Chen et al., 2024), which sets new state-of-the-art performance on the CNFOOD-241 benchmark using hybrid residual models, and SalientFusion (Song & Liu, 2025), which tackles context-aware compositional zero-shot food recognition. Additionally, recent supervised contrastive learning frameworks that leverage textual information show competitive results on Food-101 and ISIA Food-500, highlighting multimodal and curriculum strategies (Jiang et al., 2025). A comprehensive review of deep learning methods in food image recognition further contextualizes these developments (Liu et al., 2025).

2.3. *Food aesthetic*

Automatic image aesthetic assessment is a computer vision task that aims to evaluate the visual appeal of images by analyzing their adherence to aesthetic principles such as balance, harmony, contrast, or other photographic cues (Celona et al., 2021). Food images present unique challenges since the subjectivity inherent in aesthetic judgments, as perceptions of visual appeal vary widely among individuals due to personal preferences and cultural backgrounds. Moreover, domain-specific criteria unique to food imagery, like presentation style and ingredient appeal, differ from general image aesthetics and require specialized consideration (Castagna et al., 2021). Additionally, there is a notable lack of large-scale, labeled datasets specifically tailored for food image aesthetics, which hampers the development of robust assessment models.

Sheng et al. (2021, 2018) release the first large-scale dataset for assessing the aesthetics of food photographs, the Gourmet Photography Dataset (GPD). It includes 24,000 food images with human-annotated labels (aesthetically positive or negative). The authors evaluated various machine learning algorithms, demonstrating that

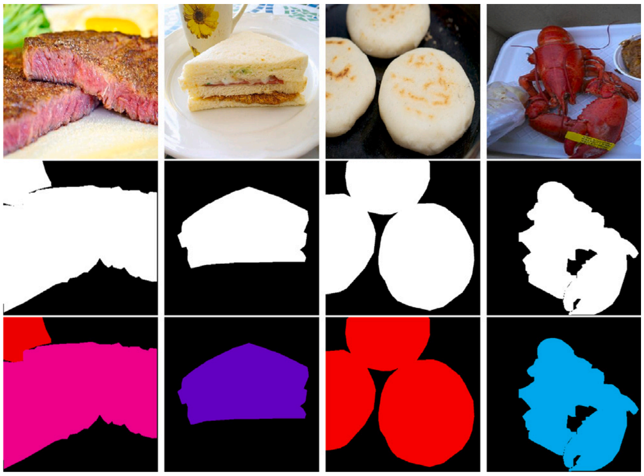


Fig. 3. Sample images in the Food-50 dataset: original images (top row); binary food/no-food groundtruth masks (center row); semantic segmentation groundtruth masks (bottom row).

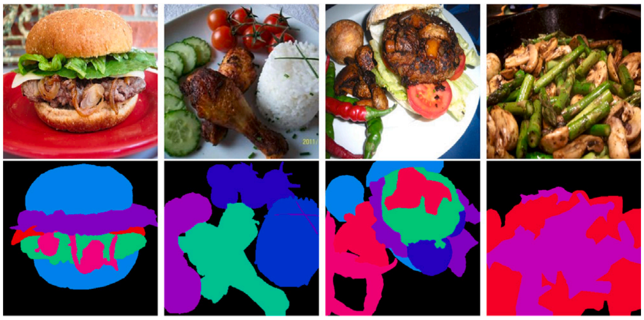


Fig. 4. Sample images in the FoodSeg103 dataset: original images (top row); semantic segmentation groundtruth masks (bottom row).

deep convolutional neural networks trained on GPD can perform on par with human experts.

Gambetti and Han (2022) propose a food aesthetics assessment model using computer vision and deep learning to estimate aesthetic scores for food images. The model utilizes a multi-stage approach, leveraging pre-trained parameters from a general aesthetics dataset and fine-tuning with a food aesthetics dataset. It is applied to social media food images, with validation conducted through human ratings via Prolific and an analysis of photographic attributes like color and composition.

Also placement of the food on plates influence the perception of food. In this regards, Zhang et al. (2022) performed several experiments, and found that food placed on more beautiful plates was perceived as tastier and healthier, while food on less beautiful plates triggered negative emotions. Food placed in the center of the plate was perceived as tastier than food at the edge.

3. Experimental setup

3.1. Datasets and evaluation metrics

In this section we describe the four datasets considered for our experiments.

Food-50 (Chen et al., 2012) contains 50 categories of worldwide food, with each category containing 100 photographs from different sources (e.g., manually taken or from internet web albums) for a total of 5000 images. The dataset is divided with stratified sampling into training and test according to a 70–30 split. Although differently stated in the paper, this is the split used. In addition to the food photographs, binary masks for food/no-food segmentation, and masks for food semantic segmentation are also available. Some sample images are reported in Fig. 3.

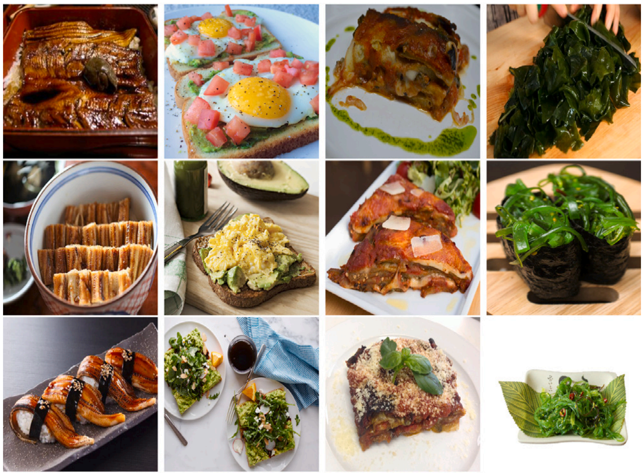


Fig. 5. Sample images in the ISIA Food-500 dataset. Left to right: samples of the class anago, avocado toast, parmigiana, and wakame.

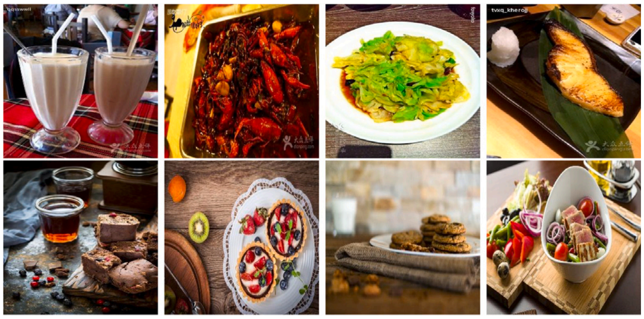


Fig. 6. Sample images in the Gourmet Photography Dataset (GPD) dataset: aesthetically negative images (top row); aesthetically positive images (bottom row).

FoodSeg103 (Wu et al., 2021) is a large scale dataset for food segmentation. It contains a total of 9490 images annotated with 103 ingredient classes (and one additional background class) with pixel-wise masks. On average each image has 6 different ingredient labels. The dataset is divided with stratified sampling into training and test according to a 70–30 split. Some sample images are reported in Fig. 4.

ISIA Food-500 (Min et al., 2020) is a large scale dataset for food recognition, containing 399,736 images belonging to 500 food categories. The dataset is divided with stratified sampling into training, validation and test according to a 50-10-30 split. Some sample images are reported in Fig. 5.

Gourmet Photography Dataset (GPD) (Sheng et al., 2018) is a dataset for the aesthetic visual assessment of food images. It contains a total of 24,000 images retrieved from various social media websites with diverse food classes and geographical information, and complemented with images retrieved from many food categorization data sets. Image labeling into aesthetically positive or aesthetically negative classes was performed using Amazon's Mechanical Turk and refined with additional expert photographers with adequate aesthetic perception. The dataset is divided with stratified sampling into training and test according to a 80–20 split. Some sample images are reported in Fig. 6.

Each task is evaluated using standard metrics commonly used by the corresponding literature. Top-1 accuracy is used for food classification,

Table 1
Summary of the different recognition tasks considered in this paper.

Task	Visual task type	Dataset(s)
Food classification	Image level	Food-50, ISIA Food-500
Food/no-food segmentation	Pixel level	Food-50
Food semantic segmentation	Pixel level	Food-50, FoodSeg103
Cross-dataset food classification	Image level	ISIA Food-500
Food aesthetic assessment	Image level	GPD
Food classification (continual learning)	Image level	Food-50, ISIA Food-500
Food classification (robustness against image distortions)	Image level	Food-50, ISIA Food-500

cross-dataset recognition, and food aesthetics assessment measuring the percentage of correctly predicted image labels. For segmentation tasks, Top-1 accuracy is used to measure the percentage of correctly predicted pixel labels. In addition, for the food semantic segmentation task, Mean Intersection over Union (mIoU) is used to assess region-level overlap between predictions and ground truth. For continual learning, performance is reported as Top-1 accuracy over all classes observed up to each incremental step.

3.2. Self-supervised features

The aim of this paper is to assess the performance of self-supervised features on a set of different food recognition tasks. The features are extracted from DINOv2 (Oquab et al., 2023), which is a foundational model trained in a self-supervised learning setting to learn general-purpose visual features both at the image and patch level. From an architectural point of view DINOv2 is a Vision Transformer ViT/14 model (Dosovitskiy, 2020) with 1B parameters then distilled (Gou et al., 2021) into a series of smaller models:

- DINOv2 ViT-S producing 384-dimensional features;
- DINOv2 ViT-B producing 768-dimensional features;
- DINOv2 ViT-L producing 1024-dimensional features;
- DINOv2 ViT-g producing 1536-dimensional features.

Concerning the patch level features, DINOv2 considers image patches with size 14×14, and thus for a 224×224 input image it produces a $16 \times 16 \times d$ feature map with d depending on the DINOv2 model size (i.e., S, B, L, or g).

In all our experiments the input images are normalized using the default ImageNet mean and standard deviation, without any data augmentation. Both the global and local patch-level features are extracted from the pre-trained DINOv2 model without any fine-tuning of the backbone and are L2-normalized to have a unitary norm before being fed to the linear classification or segmentation heads; no additional feature standardization is applied. For all the tasks considered, a linear head is independently trained from scratch; being it constituted by a single layer, no activation functions are used in order to directly assess the linear separability and generalization capability of the self-supervised features, which is the standard evaluation practice in self-supervised representation learning. After all the linear heads have been trained, they can be integrated into a single unified model. A schematic representation of the training setup of the proposed unified evaluation framework is reported in Fig. 7.

For each food classification and food aesthetic assessment task, a single linear layer is trained to map the global d-dimensional features into the number of classes c of each dataset considered, having thus a size of $d \times c + c$ where the biases are included.

For food segmentation tasks, a single convolutional layer is trained to map each of the local patch-level d-dimensional features into the number of classes c of each dataset considered, thus providing a class prediction at each spatial location. The convolutional layer is designed with 1×1 filters to preserve the spatial shape of the extracted features, and therefore it has a size of $d \times 1 \times 1 \times c + c$, biases included. We can observe how the count of trainable parameters is identical to that of the food classification tasks.

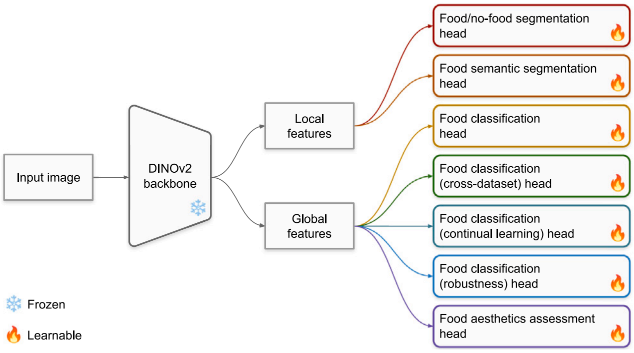

Input image
DINOv2 backbone
Local features
Global features
Food/no-food segmentation head
Food semantic segmentation head
Food classification head
Food classification (cross-dataset) head
Food classification (continual learning) head
Food classification (robustness) head
Food aesthetics assessment head
❄ Frozen
🔥 Learnable

Fig. 7. Schematic representation of the proposed unified evaluation framework for food recognition tasks.

Table 2
Top-1 accuracy for the food classification task on the Food-50 dataset.

Method	Top-1 accuracy (%)
Chen et al. (2012)	68.3
ResNet-50 (Ciocca et al., 2018a)	93.8
DINOv2 ViT-S/14 (this paper)	97.5
DINOv2 ViT-B/14 (this paper)	**98.2**
DINOv2 ViT-L/14 (this paper)	98.1
DINOv2 ViT-g/14 (this paper)	97.7

All experiments are run in PyTorch with a single Nvidia GeForce GTX 1080 gpu, using a cross entropy loss, Adam optimizer with a learning rate equal to $3 \cdot 10^{-4}$, a weight decay equal to $5 \cdot 10^{-4}$, a batch size of 16, for a total of 200 epochs.

4. Experimental results

In this section we describe the different food recognition tasks considered in order to assess the performance of DINOv2 features and compare them with the respective state of the art. In Table 1 we provide a summary of the different recognition tasks considered, providing a short description of each task, the type of visual task (i.e., if it is an image level or a pixel level task), and the dataset(s) used.

4.1. Food classification on food-50

The first experiment consists in the food classification task on the Food-50 dataset. For the DINOv2 features, a linear head is trained to classify the 50 classes. The results in terms of top-1 accuracy are reported in Table 2. From the results it is possible to see that DINOv2 features with a linear head are able to outperform the state of the art, i.e. a ResNet-50 fully trained on the same dataset, by a factor in the range [3.7%, 4.4%] depending on the size of the model considered.

Table 3
Top-1 accuracy for the food classification task on the ISIA Food-500 dataset.

Method	Top-1 accuracy (%)
VGG-16 (Min et al., 2020)	55.2
GoogLeNet (Min et al., 2020)	56.0
ResNet-152 (Min et al., 2020)	57.0
WRN-50 (Min et al., 2020)	60.1
DenseNet-161(Min et al., 2020)	60.1
NAS-NET (Min et al., 2020)	60.7
SE-ResNeXt101_32 × 4d (Min et al., 2020)	62.0
NTS-NET (Min et al., 2020)	63.7
WS-DAN (Min et al., 2020)	60.7
DCL (Min et al., 2020)	64.1
SENet-154 (Min et al., 2020)	63.8
SGLANet (Min et al., 2020)	**64.7**
DINOv2 ViT-S/14 (this paper)	54.5
DINOv2 ViT-B/14 (this paper)	60.5
DINOv2 ViT-L/14 (this paper)	63.0
DINOv2 ViT-g/14 (this paper)	62.6

4.2. Food classification on ISIA food-500

This experiment consists in the food classification task on the ISIA Food-500 dataset. For the DINOv2 features, a linear head is trained to classify the 500 classes. The results in terms of top-1 accuracy are reported in Table 3. From the results it is possible to see that DINOv2 features with a linear head are not able to outperform the state of the art, which is constituted by a neural network fully trained on the dataset. The best methods in the literature achieves an accuracy of 64.7% while the best DINOv2 features (i.e. DINOv2 ViT-L/14) achieve 63%. This behavior is expected, as ISIA Food-500 represents a closed-world, in-dataset classification scenario with a large number of fine-grained classes and sufficient labeled data. In such settings, end-to-end supervised training allows models to specialize on dataset-specific visual cues, which a frozen, general-purpose representation cannot fully exploit. DINOv2 features, by contrast, are optimized for broad semantic generalization rather than fine-grained class separation within a single dataset. As shown in subsequent experiments, this trade-off favors DINOv2 in scenarios involving domain shift or limited supervision, while task-specific models retain an advantage when high-quality labeled data is available.

4.3. Food/no-food segmentation on food-50

This experiment consists into the food/no-food segmentation on the Food-50 dataset. For the DINOv2 features a linear head is trained to segment the two classes (i.e., food and no-food). The DINOv2 features are upsampled with bilinear interpolation to the size of the groundtruth and followed by thresholding before the computation of the results. Since DINOv2 produces a feature map of size 16 × 16 for a 224 × 224 input image, we also consider the results giving a 448 × 448 input image, which results into a 32 × 32 feature map. The results in terms of top-1 accuracy are reported in Table 4. From the results it is possible to notice that DINOv2 features computed on 224 × 224 inputs are not able to outperform the state of the art, which outputs a higher resolution segmentation. Giving instead as input a 448 × 448 image, DINOv2 features outperform the state of the art by up to 0.5%.

4.4. Food semantic segmentation on food-50

This experiment consists in the food semantic segmentation on the Food-50 dataset. For the DINOv2 features a linear head is trained to segment the different food classes. The DINOv2 features are upsampled with nearest neighbor interpolation to the size of the groundtruth before the computation of the results. As for the previous experiment, we also consider the results when a 448 × 448 image is given as input to DINOv2, which results into a 32 × 32 feature map. The results in terms of top-1 accuracy are reported in Table 5. From the results it is possible to notice that DINOv2 features computed on 224 × 224 inputs are able to obtain a result close to the state of the art, from 0.2% less to 0.4% more. Giving instead as input a 448 × 448 image, DINOv2 features outperform the state of the art by up to 1.9%.

Table 4
Top-1 accuracy for the food (/no-food) segmentation task on the Food-50 dataset.

Method	Top-1 accuracy (%)
GUN (Ciocca et al., 2018a)	95.3
SSNet (Ciocca et al., 2018a)	94.9
SegNet (Ciocca et al., 2018a)	94.7
ENet (Ciocca et al., 2018a)	94.6
ERFNet (Ciocca et al., 2018a)	94.6
Edanet (Ciocca et al., 2018a)	94.6
DINOv2 ViT-S/14 (this paper)	94.3
DINOv2 ViT-B/14 (this paper)	94.3
DINOv2 ViT-L/14 (this paper)	94.1
DINOv2 ViT-g/14 (this paper)	94.4
DINOv2 ViT-S/14 (@448) (this paper)	95.7
DINOv2 ViT-B/14 (@448) (this paper)	**95.8**

Table 5
Top-1 accuracy for the food semantic segmentation task on the Food-50 dataset.

Method	Top-1 accuracy (%)
GUN (Ciocca et al., 2018a)	89.0
SSNet (Ciocca et al., 2018a)	84.6
SegNet (Ciocca et al., 2018a)	66.6
ENet (Ciocca et al., 2018a)	69.8
ERFNet (Ciocca et al., 2018a)	71.2
Edanet (Ciocca et al., 2018a)	83.5
DINOv2 ViT-S/14 (this paper)	89.0
DINOv2 ViT-B/14 (this paper)	88.9
DINOv2 ViT-L/14 (this paper)	89.4
DINOv2 ViT-g/14 (this paper)	88.8
DINOv2 ViT-S/14 (@448) (this paper)	**91.9**
DINOv2 ViT-B/14 (@448) (this paper)	90.8

4.5. Food semantic segmentation on FoodSeg103

This experiment consists in the food semantic segmentation on the FoodSeg103 dataset. Therefore the task is the same as the previous experiment, but with a dataset containing more than double the classes. The setup is identical to the one in the previous experiment with the addition of a variant of DINOv2 coupled with FeatUp (Fu et al., 2024), a model-agnostic framework for upsampling features, which is used to upscale DINOv2 features to 256 × 256 at the cost of about 200k additional model parameters.

The results in terms of top-1 accuracy and intersection over union (IoU) are reported in Table 6. From the results, it is possible to notice that DINOv2 features computed on 224 × 224 inputs obtain a result at best 18% worse than the state of the art. This is due to the high resolution of the groundtruth compared to the low-resolution features. Upsampling the DINOv2 features with FeatUp reduces this gap to about 9%; giving instead as input a 448 × 448 image, DINOv2 features reduce this gap to just 3%. These results indicate that while DINOv2 features provide strong semantic representations, their relatively low spatial resolution limits the segmentation performance in scenarios where the ground truth contains many small objects and fine-grained details (also see Fig. 4). In FoodSeg103, this leads to reduced accuracy when using standard 224 × 224 inputs, since small ingredient regions cannot be precisely localized from coarse patch-level features. Increasing the input resolution or applying feature upsampling substantially alleviates this issue, confirming that the observed limitation is mainly due to

Table 6
Top-1 accuracy for the food semantic segmentation task on the FoodSeg103 dataset.

Method	Top-1 acc. (%)	IoU
ReLeM-ViT-16/B (MLA) (Wu et al., 2021)	**57.4**	**0.451**
DINOv2 ViT-S/14 (this paper)	33.9	0.250
DINOv2 ViT-B/14 (this paper)	39.2	0.277
DINOv2 ViT-S/14 + FeatUp (this paper)	48.9	0.360
DINOv2 ViT-S/14 (@448) (this paper)	51.1	0.396
DINOv2 ViT-B/14 (@448) (this paper)	54.3	0.431

spatial resolution constraints. Upsampling necessarily comes at an additional computational cost. Let us consider DINOv2 at 224 × 224 our baseline computational cost; the computational cost of self-attention, the most expensive operation in ViTs, scales as $\mathcal{O}(T^2 \cdot d)$ with T being the number of tokens which is 256 in our case. FeatUp operates after feature extraction taking the $16 \times 16 \times d$ feature map and upsampling it to $256 \times 256 \times d$ using lightweight convolutional layers. DINOv2 still processes 256 tokens and therefore the expensive self-attention cost is unchanged, with the extra cost coming only from convolutions on feature maps. The FeatUp computational overhead is linear in spatial size and negligible compared to self-attention (less than 10% overhead relative to backbone). Using native DINOv2 upsampling, i.e. giving 448 × 448 increases the number of tokens to $T = 1024$, resulting in a self-attention which is 16 times more expensive.

4.6. Cross-dataset food classification on ISIA food-500

This experiment considers the features extracted from several neural networks trained on different food datasets, and trains on top of them a linear head to classify the ISIA Food-500 dataset, i.e. in a cross-dataset scenario. This experiment is more fair than the previous one in Section 4.2, since now we are comparing DINOv2 features with other neural features, without having neural models fully trained on the target dataset. The results in terms of top-1 accuracy are reported in Table 7. It is possible to see that DINOv2 features with a linear head are able to outperform the other features by a large margin, i.e. up to 19.4%. This cross-dataset classification experiment represents a particularly relevant real-world scenario, where a recognition system must operate on data distributions that differ from those seen during training. Unlike end-to-end supervised models, which require retraining or fine-tuning on the target dataset, DINOv2 features demonstrate strong generalization across food datasets without any adaptation of the backbone. The strong performance of DINOv2 in cross-dataset classification can be attributed to the nature of its self-supervised training, which promotes the learning of dataset-agnostic and semantically rich visual representations. Unlike supervised food recognition models, which tend to overfit to dataset-specific visual and contextual cues, DINOv2 features capture higher-level food appearance characteristics that generalize well across domains. This results in a substantial advantage when the target dataset differs from the training distribution, a scenario that closely reflects real-world deployment conditions where annotated data for new food domains may be scarce or unavailable.

4.7. Food aesthetic assessment on GPD

This experiment consists in the food aesthetic assessment task on the GPD dataset. For this experiment we extract the DINOv2 features and on top of them we train a linear head to classify the GPD dataset. The results in terms of top-1 accuracy are reported in Table 8. From the results it is possible to see that DINOv2 features with a linear head are able to obtain a performance close to the state of the art, which consists in a neural network fully trained on the GPD dataset.

4.8. Continual learning for food classification

In this experiment we want to analyze the performance of DINOv2 features for continual learning, i.e. when new data is available (Lange et al., 2021; Wang, Zhang, et al., 2024). Two scenarios are considered: in the first one, the model has to classify new classes; in the second one, the model has to classify an increasing number of classes for which an increasing number of training images is available. In all continual learning experiments, the classes are incrementally introduced following a class-incremental learning setting. At each step, a fixed number of new classes is added, while previously seen classes remain part of the test set. For Food-50, the model is initialized with 25 classes and extended in steps of 5 classes until all 50 classes are included. For ISIA Food-500, the model starts with 50 classes and is incrementally extended in steps of 50 classes all 500 classes are included. For both datasets, the classes are introduced incrementally following the lexicographic order of their labels. The DINOv2 backbone is kept frozen throughout the process, and only the classifier (linear head or k-NN) is updated. To mitigate catastrophic forgetting, Latent Replay (LR) is employed for all feature-based methods (Pellegrini et al., 2020), while Experience Replay (ER) is used for the supervised baseline (Rolnick et al., 2019). Performance is measured using top-1 classification accuracy on the test set containing all classes observed up to the current incremental step. Unless otherwise stated, reported results correspond to the accuracy after each incremental update.

Table 7
Top-1 accuracy for the cross-dataset food classification task on the ISIA Food-500 dataset.

Method	Top-1 accuracy (%)
ResNet-18 on ImageNet	26.2
ResNet-18 on Food50 @25classes	17.9
ResNet-18 on Food50 @50classes	19.8
ResNet-18 on food101 (Bossard et al., 2014)	25.2
ResNet-18 on Foodx-251 (Kaur et al., 2019)	32.9
ResNet-50 on Food-475 (Ciocca et al., 2018a)	43.6
ResNet-50 on ImageNet	29.5
ResNet-152 on ImageNet	32.0
DINOv2 ViT-S/14 (this paper)	54.5
DINOv2 ViT-B/14 (this paper)	60.5
DINOv2 ViT-L/14 (this paper)	**63.0**
DINOv2 ViT-g/14 (this paper)	62.6

Table 8
Top-1 accuracy for the food aesthetic assessment task on the GPD dataset.

Method	Top-1 accuracy (%)
Color+SVM (Sheng et al., 2018)	76.5
GIST + SVM (Sheng et al., 2018)	77.8
VGG-object + SVM (Sheng et al., 2018)	88.4
VGG-scene + SVM (Sheng et al., 2018)	85.7
VGG-food + SVM (Sheng et al., 2018)	87.7
GPD-AlexNet (Sheng et al., 2018)	77.3
GPD-VGG (Sheng et al., 2018)	88.2
GPD-InceptionV2 (Sheng et al., 2018)	89.0
GPD-ResNet (Sheng et al., 2018)	**90.8**
DINOv2 ViT-S/14 (this paper)	88.6
DINOv2 ViT-B/14 (this paper)	88.9
DINOv2 ViT-L/14 (this paper)	88.1
DINOv2 ViT-g/14 (this paper)	87.5

4.8.1. Class-incremental continual learning

In the first part of this experiment we consider the problem of class-incremental continual learning on the Food-50 dataset. We start with 25 classes and we increment them in steps of 5 until all the 50 dataset classes are covered. On top of the DINOv2 features we train both a linear head and a k−NN classifier. For comparison we also include the features extracted from a ResNet-18 trained on ImageNet, and the features of a ResNet-18 trained on the first 25 Food-50 classes considered in this experiment. Furthermore, we also include a ResNet-18 supervisely trained on each considered class cardinality using Experience Replay (ER) continual learning (Rolnick et al.,

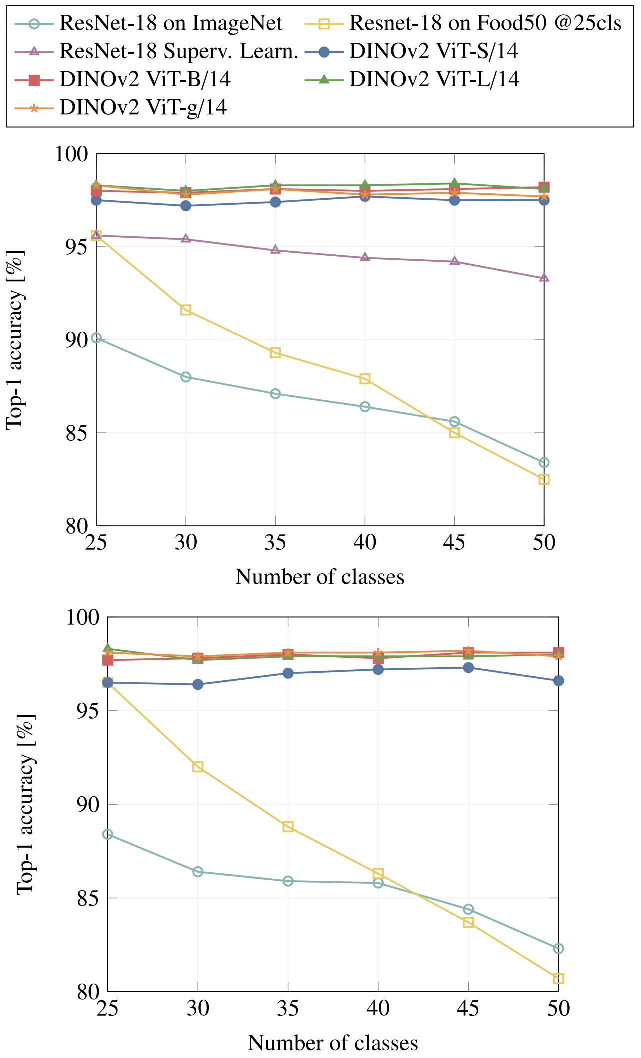


Fig. 8. Top-1 accuracy for the food classification task with class-incremental continual learning on the Food-50 dataset: linear head (top); k−NN (bottom). Accuracy is computed on the union of all classes introduced up to the current incremental step.

2019). All the other configurations are trained using Latent Replay (LR) continual learning (Pellegrini et al., 2020).

The classification results are reported in terms of top-1 accuracy in Table 9 and plotted in Fig. 8. From the results it is possible to notice that using a linear head or a k−NN classifier yield quite similar performance, especially for the larger DINOv2 models features. It is also possible to see how food-specific features learned on the same dataset with 25 classes, struggle to generalize to new classes, to the point that the test on 50 classes performs even worse than ImageNet features. On the other side, we observe how the performance of DINOv2 features remains almost constant with respect to the number of considered classes. Furthermore, we notice that DINOv2 features obtain an accuracy that is even higher than the ResNet-18 based on supervised training.

In the second part of this experiment, we consider the problem of class-incremental continual learning on the ISIA Food-500 dataset. We start with 50 classes and we increment them in steps of 50 until all the 500 classes are considered. This experiment is designed to be much harder than the previous one, since this corresponds to a tenfold increase in the number of classes. Since the classification results using a linear head and a k−NN classifier are similar, in this experiment only the latter is considered.

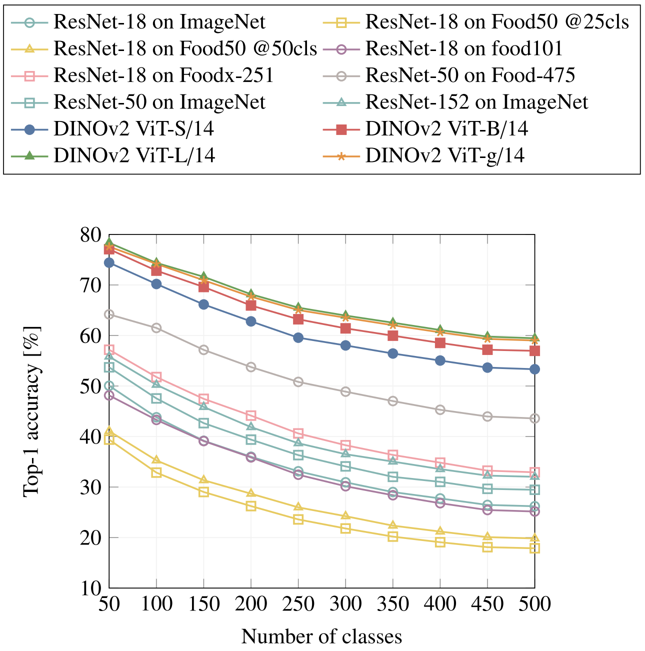


Fig. 9. Top-1 accuracy for the food classification task with class-incremental continual learning on the ISIA Food-500 dataset. Accuracy is computed on the union of all classes introduced up to the current incremental step.

Performance in terms of top-1 accuracy is plotted in Fig. 9. We can observe that, as expected, all the features exhibit a reduction in accuracy as the number of classes increases. Regarding the individual features, we can see the formation of different groups of performance levels: in the top one there are the DINOv2 features, with DINOv2-S being the worst of this group. In the bottom one there are the food-specific features trained on the smaller datasets and the ImageNet features, with ResNet-152 ImageNet features and ResNet-18 Foodx-251 features being the best ones of this group. The latter is a deep ResNet variant trained on the Foodx-251 dataset (Kaur et al., 2019). Halfway between the two groups there are the ResNet-50 food475 features: a ResNet variant trained on the Food475 dataset (Ciocca et al., 2018b). In general we observe how the best performance among competitor methods is obtained by very large features trained on a large food dataset (ResNet-50, with a feature size of 2048 trained on Food-475).

4.8.2. Incremental and class-incremental continual learning

In this experiment we consider the incremental and class-incremental continual learning task on the Food-50 dataset. The number of classes is varied as in the previous experiment (i.e., from 25 to 50 in steps of 5), while the number of training images considered is varied among $[1, 2, 3, 5, 7, 8, 10, 15, 20, 25, 30, 40, 50, 60, 70]$. For each training set cardinality, ten independent random extractions are performed. For all the features considered, a k−NN classifier is used. The performance in terms of top-1 accuracy, averaged over the ten independent runs, is plotted in Fig. 10 with one plot for each number of classes considered.

From the plots we can observe how the performance of the different features increases as the number of training images increases. We can also notice how the performance of the DINOv2 features and the features extracted from the ResNet-18 trained on ImageNet are less dependent from the number of images used for training with respect to the features extracted from the ResNet-18 trained on Food-50. Furthermore, we can see that in the most difficult case (i.e., 50 classes), DINOv2 features with just one training image per class perform equivalently (in the case of DINOvs-S) or even much better (in the case of the larger DINOv2 models) than the other features considered. In

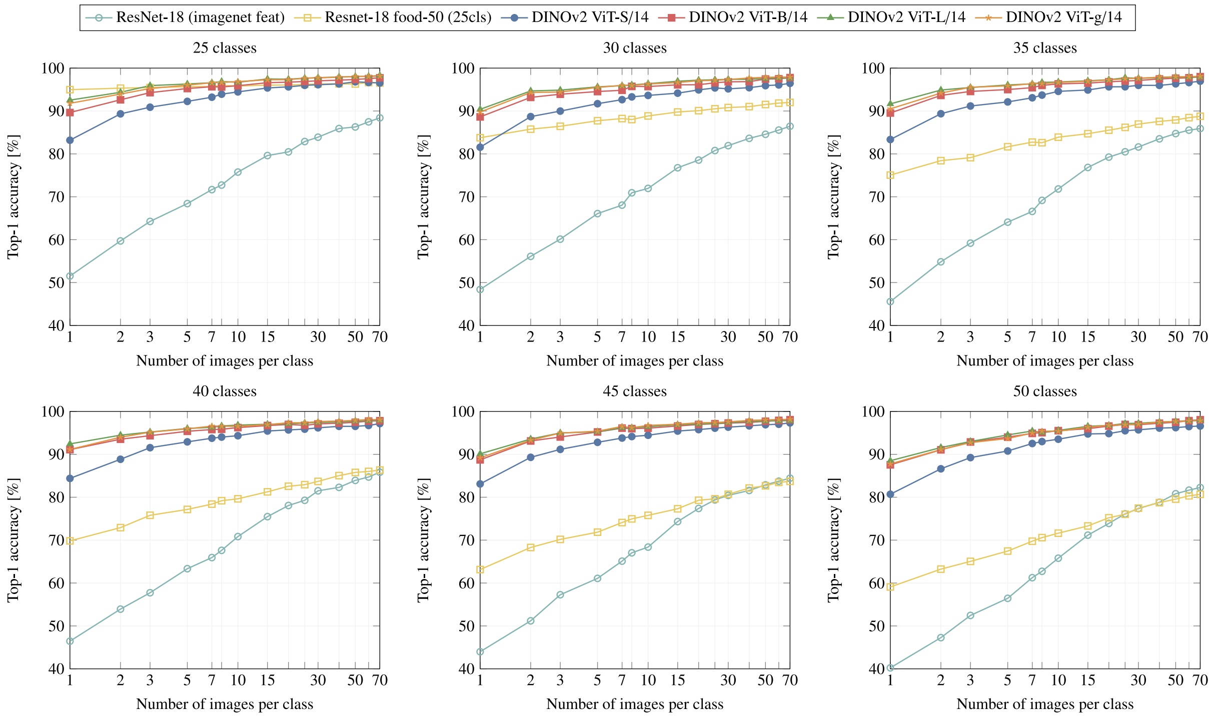


Fig. 10. Top-1 accuracy for the food classification task with incremental and class-incremental continual learning on the Food-50 dataset. Accuracy is computed on the union of all classes introduced up to the current incremental step.

Table 9
Top-1 accuracy (%) for the food classification task with class-incremental continual learning on the Food-50 dataset: (a) linear head; (b) k−NN. Accuracy is computed on the union of all classes introduced up to the current incremental step..

(a)

Method	Number of classes					
	25	30	35	40	45	50
ResNet-18 on ImageNet	90.1	88.0	87.1	86.4	85.6	83.4
Resnet-18 on Food50 @25 cls.	95.6	91.6	89.3	87.9	85.0	82.5
ResNet-18 Superv.Learn.	95.6	95.4	94.8	94.4	94.2	93.3
DINOv2 ViT-S/14 (this paper)	97.5	97.2	97.4	97.7	97.5	97.5
DINOv2 ViT-B/14 (this paper)	98.0	97.9	98.1	98.0	98.1	**98.2**
DINOv2 ViT-L/14 (this paper)	**98.3**	**98.0**	**98.3**	**98.3**	**98.4**	98.1
DINOv2 ViT-g/14 (this paper)	**98.3**	97.8	98.1	97.8	97.9	97.7

(b)

Method	Number of classes					
	25	30	35	40	45	50
ResNet-18 on ImageNet	88.4	86.4	85.9	85.8	84.4	82.3
Resnet-18 on Food50 @25 cls.	96.5	92.0	88.8	86.3	83.7	80.7
DINOv2 ViT-S/14 (this paper)	96.5	96.4	97.0	97.2	97.3	96.6
DINOv2 ViT-B/14 (this paper)	97.7	97.8	98.0	97.8	98.1	**98.1**
DINOv2 ViT-L/14 (this paper)	**98.3**	97.7	97.9	97.9	97.9	98.0
DINOv2 ViT-g/14 (this paper)	98.1	**97.9**	**98.1**	**98.1**	**98.2**	97.9

Table 10
Top-1 accuracy for the food classification task on the distorted ISIA Food-500 dataset subdivided by distortion type (left), and percentual difference in top-1 accuracy with respect to the case when no distortions are applied (right). The first column reports the performance when no image distortions are applied (No dist.); the next three columns report the results when blur (B), global illuminant change (I), or noise (N) are individually applied; the next three columns report the results when two different distortions are applied in sequence (B+I, B+N, I+N); finally, the last column report the results when all the three distortions considered are sequentially applied (B+I+N).

(a)

Method	Top-1 accuracy							
	No dist.	B	I	N	B+I	B+N	I+N	B+I+N
ResNet18 on Foodx-251	29.0	18.8	21.6	6.2	12.2	2.7	4.3	1.8
ResNet50 on Food-475	39.1	28.2	35.1	10.4	24.0	3.8	7.8	3.2
ResNet152 on ImageNet	28.1	20.9	24.7	10.3	17.5	5.2	8.2	4.0
DINOv2 ViT-S/14 (this paper)	51.9	42.9	51.1	33.1	39.1	20.3	27.9	16.1
DINOv2 ViT-B/14 (this paper)	56.2	45.8	55.3	40.0	43.6	25.7	35.8	22.6
DINOv2 ViT-L/14 (this paper)	**58.0**	49.0	**57.6**	47.2	48.6	35.5	44.1	32.8
DINOv2 ViT-g/14 (this paper)	57.7	**50.4**	57.4	**49.0**	**49.5**	**39.4**	**46.4**	**36.8**

(b)

Top-1 accuracy percentual difference wrt No dist.							
No dist.	B	I	N	B+I	B+N	I+N	B+I+N
–	−35.1	−25.4	−78.5	−58.0	−90.8	−85.0	−93.9
–	−27.8	−10.2	−73.3	−38.6	−90.4	−80.0	−91.8
–	−25.4	−12.2	−63.5	−37.5	−81.6	−70.9	−85.8
–	−17.4	−1.6	−36.4	−24.8	−60.9	−46.2	−69.0
–	−18.4	−1.5	−28.8	−22.3	−54.2	−36.2	−59.7
–	−15.5	−0.8	−18.6	−16.2	−38.9	−24.0	−43.5
–	**−12.7**	**−0.6**	**−15.2**	**−14.3**	**−31.8**	**−19.7**	**−36.2**

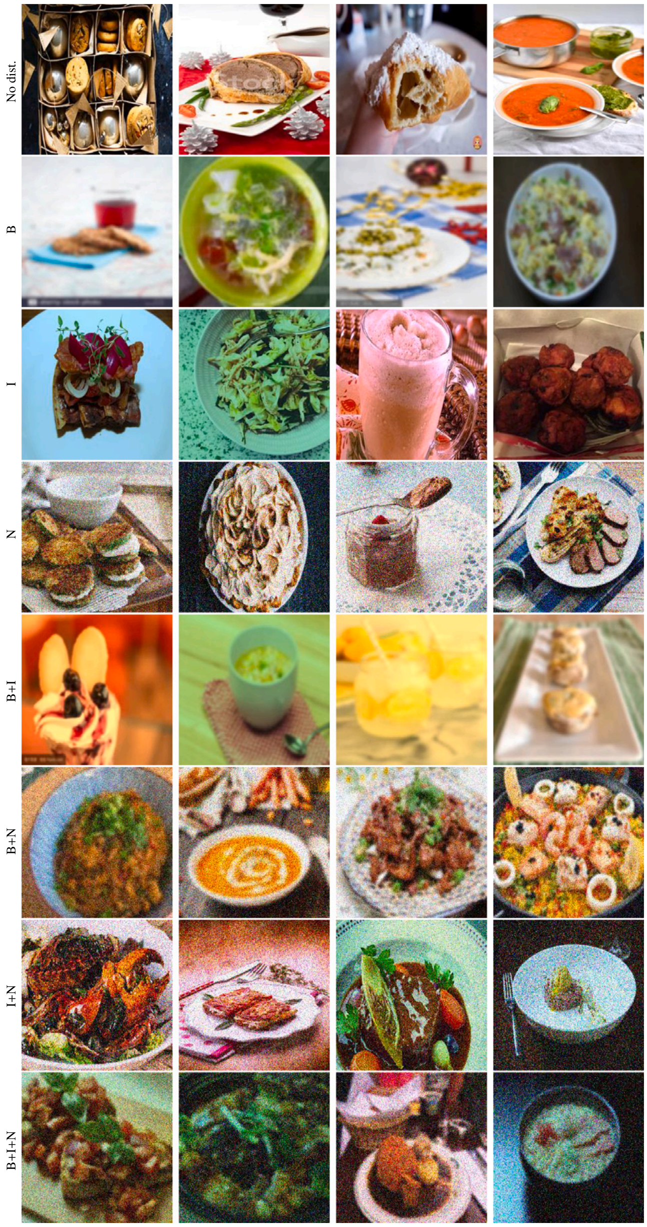


Fig. 11. Sample images in the distorted ISIA Food-500 dataset. Top to bottom: sample images when no image distortions are applied (No dist.); sample images when a single distortion is applied: blur (B), global illuminant change (I), or noise (N); sample images when two different distortions are applied in sequence (B+I, B+N, I+N); sample images when all the three distortions considered are sequentially applied (B+I+N).

particular we notice that DINOv2 features with just one training image per class perform similarly to the other features with 50 training images per class.

4.9. Robustness against image distortions

In this experiment we want to analyze the robustness of DINOv2 features against image distortions for food classification on the ISIA Food-500 dataset. In particular three different distortions are considered and applied to the test set: blur, global illuminant change, and noise. All the possible combinations of distortions are considered: no distortions, a single distortion, two distortions in sequence, and all the three distortions in sequence. For the blur distortion a 2-D Gaussian smoothing kernel with standard deviation randomly sampled with uniform distribution in the range $[0.5, 2.5]$ is considered. For the global illuminant change the diagonal Von Kries model is used with an independent scaling factor for the red, blue and green color channels; the scaling factors for the red and blue channels are randomly sampled with uniform distribution in the range $[0.5, 1.5]$, while the scaling factor for the green channel is fixed to 1 to preserve the luminance of the image. For the noise distortion we add to the image zero-mean, Gaussian white noise with a variance randomly sampled with uniform distribution in the range $[0.01, 0.05]$. Some samples of distorted images are reported in Fig. 11.

For all the considered features, a k−NN model is trained. The results in terms of top-1 accuracy are reported in Table 10. In order to better evaluate the robustness of each feature considered, the percentage difference between the top-1 accuracy of the considered distortion type, and the top-1 accuracy in absence of distortions is also reported in the same table. From the reported results it is possible to see that DINOv2 features are the ones obtaining the best performance across all the considered distortion combinations, with DINOv2-L and DINOv2-g obtaining the highest accuracy. Concerning the robustness of the features, we observe how the DINOv2 are the ones having the lowest difference with respect to the case where no distortions are applied. In particular, DINOv2-g features are the most robust against image distortions of all the considered features.

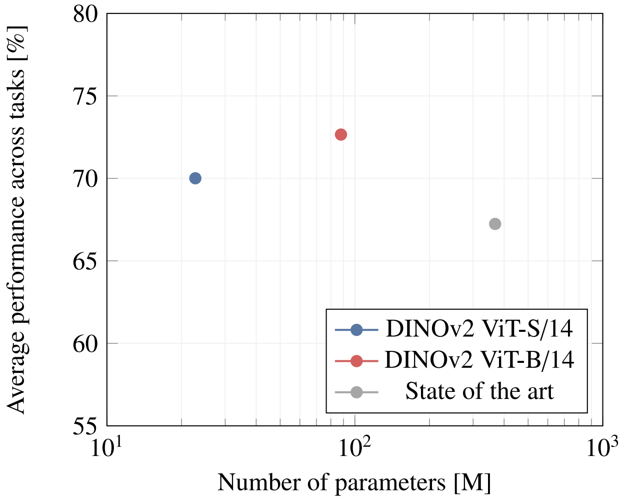
4.10. Performance summary and computational complexity

To summarize the performance on the different tasks considered we report a radar plot in Fig. 2. For the state-of-the-art performance the best method for each task is reported, while for the proposed unified framework only the features extracted from the same pre-trained model (i.e., DINOv2-B) are considered for all the tasks, although on individual tasks other DINOv2 backbones obtained better performance. DINOv2-B is chosen since it is on average it is the best performing on the tasks considered. For most of the tasks the performance is directly taken from the corresponding tables, while for continual learning we report the average accuracy over the number of classes considered on the ISIA Food-500 dataset, and for the robustness against image distortions we report the average accuracy over the single and multiple distortions considered.

In Fig. 12 instead we compare the average performance of DINOv2-B and DINOv2-S with that of the state of the art against the total number of model parameters. We can observe how both DINOv2-S and DINOv2-B are able to outperform the state of the art by about 2.8% and 5.4% respectively. Moreover, this improvement is obtained at a fraction of the number of model parameters: DINOv2-S and DINOv2-B have respectively about 6.2% and 23.9% of the parameters of the best state-of-the-art methods when all the tasks are considered. Considering the single tasks, DINOv2-S has on average 50.7% of the parameters with respect to the best state-of-the-art method, while DINOv2-B has 198.4% of the parameters.

5. Practical implications for real-world food recognition

The findings of this study have direct implications for real-world food recognition systems, where robustness, scalability, and adaptability are often more critical than peak in-dataset accuracy. In industrial quality control and manufacturing, where food items may vary across production lines or acquisition conditions, the strong cross-dataset generalization of DINOv2 features enables reliable deployment without retraining for every new setting. Similarly, in supermarket sorting and inventory systems, a unified feature representation can support multiple tasks such as classification and segmentation, while reducing the need for task-specific models. For consumer applications such as calorie and nutrient estimation, the ability of self-supervised features to generalize across diverse food appearances and cuisines is particularly valuable, as annotated data is often limited or unavailable. At the



Task Task	Number of parameters [M] DINOv2 ViT-S/14	 DINOv2 ViT-B/14	 State of the art
Backbone	22.06	86.58	–
Food classif. Food-50	0.02	0.04	25.56
Food/no-food segm. Food-50	0.02	0.04	16.40
Food semantic segm. Food-50	0.02	0.04	16.40
Food semantic segm. FoodSeg103	0.04	0.08	86.57
Food classif. ISIA Food-500	0.19	0.38	160.26
X-dataset food classif. ISIA Food-500	0.19	0.38	25.56
Food aesthetic assessment GPD	0.00	0.00	11.69
Food classif.: continual learning ISIA Food-500 (†)	0.19	0.38	25.56
Food classif.: distortion robustness ISIA Food-500 (⋆)	0.19	0.38	25.56
Total	22.73	87.92	368.01

(†) We just report the number of parameters for the 500 classes.

(⋆) Parameters not added in the total: it is an analysis performed on the model trained for classification.

Fig. 12. Top: Average performance across all the tasks considered obtained by the best state-of-the-art approaches (considering a different method for each task) and by the proposed unified framework with respect to the total number of parameters. For the proposed unified framework only the variants with a number of parameters lower than that of the state of the art are considered: DINOv2 ViT-S/14 and DINOv2 ViT-B/14. Bottom: Detailed number of parameters for each task.

same time, our results indicate that high-resolution supervision or task-specific fine-tuning remains important for applications requiring precise localization or fine-grained discrimination, highlighting a trade-off between generalization and specialization that practitioners should consider. This principle of balancing general-purpose features with targeted fine-tuning extends beyond food analysis. The empirical insights provided by this study may also inform representation-centric design choices in other AI-assisted decision-making pipelines, beyond food analysis (Ghorbani et al., 2025; Ye et al., 2025).

6. Conclusions

In this paper, we conducted a comprehensive empirical study on the use of DINOv2 self-supervised features for food recognition tasks showing that they achieve state-of-the-art or near state-of-the-art performance across multiple benchmarks. Specifically, across such datasets as Food-50, FoodSeg103, and ISIA Food-500, DINOv2 features deliver high accuracy even in challenging scenarios like semantic segmentation and cross-dataset generalization. DINOv2 features enable the simultaneous handling of tasks such as segmentation, recognition, and aesthetic assessment, thus simplifying model deployment and reducing the computational complexity for a food recognition pipeline eliminating the need for task-specific supervised models. Furthermore, DINOv2 features exhibit strong resilience to common image distortions, including blur, noise, and illumination changes, ensuring reliable performance in practical applications where image quality may vary. Finally, in a continual learning setting DINOv2 features proved their robustness achieving high recognition accuracy even with very few images per class. This highlights the potential of using them in real scenarios where data distributions and tasks evolve over time.

Although DINOv2 represents a major step forward, opportunities for further improvement include extending the evaluation to larger and more diverse food datasets, such as those covering additional cuisines, acquisition conditions, or long-tail class distributions, to further assess generalization under more challenging real-world scenarios. Another promising direction is improving inference speed and efficiency through model distillation, feature compression, or lightweight adaptation strategies, enabling deployment for real-time recognition on mobile and edge devices where computational resources are limited.

CRediT authorship contribution statement

Simone Bianco: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Marco Buzzelli:** Methodology, Validation, Visualization, Writing – original draft. **Gianluigi Ciocca:** Validation, Data curation, Writing – review & editing. **Flavio Piccoli:** Methodology, Software, Investigation, Writing – original draft. **Raimondo Schettini:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title "ON Foods - Research and innovation network on food and nutrition Sustainability, Safety and Security – Working ON Foods".

This work was partially supported by the MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Data availability

Data will be made available on request.

References

Allegra, D., Battiato, S., Ortis, A., Urso, S., & Polosa, R. (2020). A review on food recognition technology for health applications. *Health Psychology Research*, *8*(3).

Aslan, S., Ciocca, G., Mazzini, D., & Schettini, R. (2020). Benchmarking algorithms for food localization and semantic segmentation. *International Journal of Machine Learning and Cybernetics*, *11*, 2827–2847.

Bianco, S., Buzzelli, M., Chiriaco, G., Napoletano, P., & Piccoli, F. (2023). Food recognition with visual transformers. In *2023 IEEE 13th international conference on consumer electronics-berlin (ICCE-berlin)* (pp. 82–87). IEEE.

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13* (pp. 446–461). Springer.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650–9660).

Castagna, A. C., Pinto, D. C., Mattila, A., & de Barcellos, M. D. (2021). Beauty-is-good, ugly-is-risky: Food aesthetics bias and construal level. *Journal of Business Research*, *135*, 633–643.

Celona, L., Ciocca, G., & Napoletano, P. (2021). A grid anchor based cropping approach exploiting image aesthetics, geometric composition, and semantics. *Expert Systems with Applications*, *186*, Article 115852.

Chen, C.-S., Chen, G.-Y., Zhou, D., Jiang, D., & Chen, D.-S. (2024). Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning. arXiv preprint arXiv:2402.15761.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PmLR.

Chen, M.-Y., Yang, Y.-H., Ho, C.-J., Wang, S.-H., Liu, S.-M., Chang, E., Yeh, C.-H., & Ouhyoung, M. (2012). Automatic chinese food identification and quantity estimation. In *SIGGRAPH Asia 2012 technical briefs* (pp. 1–4).

Ciocca, G., Napoletano, P., & Schettini, R. (2017). Food recognition: a new dataset, experiments and results. *IEEE Journal of Biomedical and Health Informatics*, *21*(3), 588–598.

Ciocca, G., Napoletano, P., & Schettini, R. (2018a). CNN-based features for retrieval and classification of food images. *Computer Vision and Image Understanding*, *176*, 70–77.

Ciocca, G., Napoletano, P., & Schettini, R. (2018b). CNN-based features for retrieval and classification of food images. *Computer Vision and Image Understanding*, *176–177*, 70–77.

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3366–3385.

Dehais, J., Anthimopoulos, M., & Mougiakakou, S. (2016). Food image segmentation for dietary assessment. In *Proceedings of the 2nd international workshop on multimedia assisted dietary management* (pp. 23–28).

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Ege, T., Shimoda, W., & Yanai, K. (2019). A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice. In *Proceedings of the 5th international workshop on multimedia assisted dietary management* (pp. 82–87).

Fakhrou, A., Kunhoth, J., & Al Maadeed, S. (2021). Smartphone-based food recognition system using multiple deep cnn models. *Multimedia Tools and Applications*, *80*(21), 33011–33032.

Fu, S., Hamilton, M., Brandt, L., Feldman, A., Zhang, Z., & Freeman, W. T. (2024). Featup: A model-agnostic framework for features at any resolution. arXiv preprint arXiv:2403.10516.

Gambetti, A., & Han, Q. (2022). Camera eats first: exploring food aesthetics portrayed on social media using deep learning. *International Journal of Contemporary Hospitality Management*, *34*(9), 3300–3331.

Ghorbani, H., Papyan, H., Minasyan, A., Wood, D. A., Ghorbani, P., Ghorbani, S., Avagyan, E., Badakian, S., & Minasian, N. (2025). A robust multi-criteria decision-making approach for selecting optimal drugs in epilepsy treatment using the analytic hierarchy process. *Brain Disorders*, Article 100292.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789–1819.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, *33*, 21271–21284.

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 9052–9071.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).

Honbu, Y., & Yanai, K. (2022). Unseen food segmentation. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 19–23).

Jiang, S., Min, W., Liu, L., & Luo, Z. (2019). Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing*, *29*, 265–276.

Jiang, F., Ye, Z., Zhou, L., & Huang, J. (2025). Text enhanced curriculum supervised contrastive learning for food image recognition. *Neurocomputing*, Article 131781.

Kagaya, H., Aizawa, K., & Ogawa, M. (2014). Food detection and recognition using convolutional neural network. In *Proceedings of the 22nd ACM international conference on multimedia* (pp. 1085–1088).

Kaur, P., Sikka, K., Wang, W., Belongie, S., & Divakaran, A. (2019). Foodx-251: a dataset for fine-grained food classification. arXiv preprint arXiv:1907.06167.

Kawano, Y., & Yanai, K. (2015). Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, *74*, 5263–5287.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026).

Kong, N. A., Moy, F. M., Ong, S. H., Tahir, G. A., & Loo, C. K. (2023). MyDietCam: Development and usability study of a food recognition integrated dietary monitoring smartphone application. *Digital Health*, *9*, Article 20552076221149320.

Lan, X., Lyu, J., Jiang, H., Dong, K., Niu, Z., Zhang, Y., & Xue, J. (2023). Foodsam: Any food segmentation. *IEEE Transactions on Multimedia*.

Liu, L., Guan, Y., Wang, Z., Shen, R., Zheng, G., Fu, X., Yu, X., & Jiang, J. (2024). An interactive food recommendation system using reinforcement learning. *Expert Systems with Applications*, Article 124313.

Liu, C., Liang, Y., Xue, Y., Qian, X., & Fu, J. (2020). Food and ingredient joint learning for fine-grained recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(6), 2480–2493.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 857–876.

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.

Marın, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., & Torralba, A. (2021). Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 187–203.

Mezgec, S., & Koroušić Seljak, B. (2017). NutriNet: a deep learning food and drink image recognition system for dietary assessment. *Nutrients*, *9*(7), 657.

Min, W., Liu, L., Wang, Z., Luo, Z., Wei, X., Wei, X., & Jiang, S. (2020). 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* (pp. 393–401).

Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., & Jiang, S. (2023). Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(8), 9932–9949.

Nguyen, H.-T., Cao, Y., Ngo, C.-W., & Chan, W.-K. (2024). FoodMask: Real-time food instance counting, segmentation and recognition. *Pattern Recognition*, *146*, Article 110017.

Okamoto, K., & Yanai, K. (2021). UEC-foodpix complete: A large-scale food image segmentation dataset. In *Pattern recognition. ICPR international workshops and challenges: virtual event, January 10–15, 2021, proceedings, part v* (pp. 647–659). Springer.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Pellegrini, L., Graffieti, G., Lomonaco, V., & Maltoni, D. (2020). Latent replay for real-time continual learning. In *2020 IEEE/RSJ international conference on intelligent robots and systems* (pp. 10203–10209). IEEE.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., & Wayne, G. (2019). Experience replay for continual learning. *Advances in Neural Information Processing Systems*, *32*.

Sheng, K., Dong, W., Huang, H., Chai, M., Zhang, Y., Ma, C., & Hu, B.-G. (2021). Learning to assess visual aesthetics of food images. *Computational Visual Media*, *7*, 139–152.

Sheng, K., Dong, W., Huang, H., Ma, C., & Hu, B.-G. (2018). Gourmet photography dataset for aesthetic assessment of food images. In *SIGGRAPH Asia 2018 technical briefs* (pp. 1–4).

Shukor, M., Couairon, G., Grechka, A., & Cord, M. (2022). Transformer decoders with multimodal regularization for cross-modal food retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4567–4578).

Sinha, G., Parmar, K., Azimi, H., Tai, A., Chen, Y., Wong, A., & Xi, P. (2023). Transferring knowledge for food image segmentation using transformers and convolutions. arXiv preprint arXiv:2306.09203.

Song, J., & Liu, X. (2025). SalientFusion: Context-aware compositional zero-shot food recognition. In *International conference on artificial neural networks* (pp. 236–248). Springer.

Vlachopoulou, V., Sarafis, I., & Papadopoulos, A. (2023). Food image classification and segmentation with attention-based multiple instance learning. In *18th international workshop on semantic and social media adaptation & personalization (SMAP 2023)* (pp. 1–5).

Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5362–5383.

Wang, J., Zheng, Y., Wang, J., Xiao, X., Sun, J., & Hou, S. (2024). RD-FGM: A novel model for high-quality and diverse food image generation and ingredient classification. *Expert Systems with Applications*, Article 124720.

Wu, X., Fu, X., Liu, Y., Lim, E.-P., Hoi, S. C., & Sun, Q. (2021). A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 506–515).

Ye, L., Li, X., Ghorbani, H., Zhou, X., Minasyan, A., Zheng, B., Pan, X., Zhang, G., & Chen, D. (2025). Artificial intelligence in rheumatology: A transformative perspective. *Journal of Translational Internal Medicine*, *13*(5), 390–393.

Zhang, S., Qian, J., Wu, C., He, D., Zhang, W., Yan, J., & He, X. (2022). Tasting more than just food: Effect of aesthetic appeal of plate patterns on food perception. *Foods*, *11*(7), 931.